Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

On the nonparametric prediction of conditionally stationary sequences

S. Caires, J.A. Ferreira

REPORT PNA-R0304 MAY 31, 2003

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

On the Nonparametric Prediction of Conditionally Stationary Sequences

S. Caires
KNMI, Royal Netherlands Meteorological Institute
P.O. Box 201, NL-3730 AE De Bilt, The Netherlands
caires@knmi.nl

J.A. Ferreira
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
jose.ferreira@cwi.nl

ABSTRACT

We prove the strong consistency of estimators of the conditional distribution function and conditional expectation of a future observation of a discrete time stochastic process given a fixed number of past observations. The results apply to conditionally stationary processes (a class of processes including Markov and stationary processes) satisfying a strong mixing condition, and they extend and bring together the work of several authors in the area of nonparametric estimation. One of our goals is to provide further justification for the growing practical application of estimators in non-stationary time series and in other 'non i.i.d.' settings. Some arguments as to why such estimators should work very generally in practice, often in a nearly 'optimal' way, are given. Two numerical illustrations are included, one with simulated data and the other with oceanographic data.

2000 Mathematics Subject Classification: 62G08, 62G30, 62G07, 62G15, 62-07.
Keywords and Phrases: Nonparametric prediction, conditional distribution function, conditional expectation, time series, data analysis.
Note: The research of the first author was funded by the European Commission ERA-40 Project (no. EVK2-CT-1999-00027). The work of the second author was carried out under the project PNA3.3, 'Stochastic Processes and Applications', and funded by The Fifth Framework Programme of the European Commission through the Dynstoch research network.

1 Introduction

Let $X = \{X_i : i \in \mathbb{N}\}$ be a sequence of real-valued random variables defined on a probability space $(\Omega, \mathcal{F}, P)$. We consider the problem of predicting the value of X_{n+1} given only the knowledge of the past observations $X_n, \ldots, X_1$ when little is known about the distribution of X. Solutions to this problem include an estimate of the conditional distribution function of X_{n+1} given $(X_n, \ldots, X_1) = (x_n, \ldots, x_1) \in \mathbb{R}^n$,

$$P\big(X_{n+1} \le x \big| X_n = x_n, \ldots, X_1 = x_1\big), \quad x \in \mathbb{R}, \tag{1.1}$$

an estimate of the conditional mean $E[X_{n+1}|X_n = x_n, \ldots, X_1 = x_1]$ (when this exists), an estimate of the median of (1.1), and a prediction interval $\big(x^{(n)}_{\epsilon/2}, x^{(n)}_{1-\epsilon/2}\big)$ such that

$$P\left(X_{n+1} \in \big(x^{(n)}_{\epsilon/2}, x^{(n)}_{1-\epsilon/2}\big) \middle| X_n = x_n, \ldots, X_1 = x_1\right) \ge 1 - \epsilon$$

for a given $\epsilon \in (0,1)$.

Since prediction is to be carried out solely on the basis of the present realization of the sequence, any approach to the problem must consist of treating, in one way or another, the realization of the process as samples of the process itself. One very general form of this approach, which is being more and more frequently applied in practical problems, is based on a simple and intuitive idea pervading all those areas and sub-areas of science sometimes put under the heading of 'statistical learning', such as neural networks and nearest-neighbour methods: In order to predict the *outcome of an event* (a 'response' variable, say) under a particular *context* (a collection of measurements, initial conditions, 'explanatory variables', 'features', etc.), we may look back into our *past history* (a sample or 'training set') for situations where the same or approximately the same context was observed, and (if at least one such instance is found) predict the outcome of the event in question on the basis of what the homologous outcomes were in the past, for example by averaging them or by choosing the most frequent among them.

In the context of our time series X the most natural way of implementing this idea is perhaps to fix a positive integer $m < n$, construct estimators of the conditional distribution function, expectation or quantiles of X_{n+1} given the previous m observations $X_n, \ldots, X_{n-m+1}$, and use these as tools to make inferences and predictive statements about the future observation. The main objective of this paper is to show that a particular class of such estimators is consistent under rather general assumptions, a conclusion which will have at least the virtue of encouraging and justifying even more their application in practical problems.

For the idea outlined above to work it seems necessary, at least from a technical point of view, to assume that X is *conditionally stationary* in the sense that the conditional distribution function of X_{n+1} given $(X_n, \ldots, X_{n-m+1}) = u \in \mathbb{R}^m$ does not depend on n. Accordingly, we shall assume in everything that follows that there exists a so-called probability kernel $(u, v) \to F(v|u)$ such that

$$\int_{[U_i \in B]} F(v|U_i) dP = \int_{[U_n \in B]} P\left(X_{n+1} \leq v | X_n, \ldots, X_{n-m+1}\right) dP = \int_{[U_m \in B]} P\left(V_m \leq v | U_m\right) dP$$

for all $v \in \mathbb{R}$ and $B \in \mathcal{B}^m$, where

$$U_i = \left(X_i, \ldots, X_{i-m+1}\right), \quad V_i = X_{i+1}, \quad i \geq m. \tag{1.2}$$

For simplicity we shall avoid indicating the dependence on m in our notation for $F(\,\cdot\,|u)$, but since m will always have the same meaning throughout the paper this should not be a source of confusion.

If $F(\,\cdot\,|u)$ has a first moment for u in a given set, we shall call $u \to R(u) = \int v dF(v|u) \equiv E\left[V_i | U_i = u\right]$ the *regression function* (of V_i on U_i).

Write $S(u,h) = \{u' \in \mathbb{R}^m : -h/2 < u_i - u_i' < h/2,\, i = 1, \ldots, m\}$ for the m-dimensional square centred at u with sides of length $h > 0$ parallel to the coordinate axes. Given a sequence $\{h_n\}$ of strictly positive numbers converging to 0, we define the *empirical conditional distribution function* (of X_{n+1} given $(X_n, \ldots, X_{n-m+1}) = u$) based on (1.2) by

$$F_n(v|u) = \frac{\sum_{i=m}^{n-1} 1_{[V_i \leq v, U_i \in S(u,h_n)]}}{\sum_{i=m}^{n-1} 1_{[U_i \in S(u,h_n)]}}, \quad v \in \mathbb{R}, \tag{1.3}$$

and the *empirical regression function* by

$$R_n(u) = \int v \, dF_n(v|u) = \frac{\sum_{i=m}^{n-1} V_i 1_{[U_i \in S(u,h_n)]}}{\sum_{i=m}^{n-1} 1_{[U_i \in S(u,h_n)]}}. \tag{1.4}$$

These functions are defined only on the set $\left\{ \sum_{i=m}^{n-1} 1_{[U_i \in S(u,h_n)]} > 0 \right\}$; they may be arbitrarily defined elsewhere. From a more general standpoint we shall regard them as *regression estimators* based on an arbitrary sequence $(U_1, V_1), (U_2, V_2), \ldots$ which in particular need not have anything to do with the time series $X_1, X_2, \ldots$ Then (1.3) and (1.4) are essentially regression estimators of the Nadaraya-Watson type (Nadaraya (1964), Watson (1964)), studied in various forms and under different contexts by many authors. Stute (1986) considered the conditional empirical distribution function (1.3) in the i.i.d. case, proving its a.s. uniform convergence to $F(\,\cdot\,|u)$ for a.a. u solely under the condition that $nh_n^m / \log n \to \infty$.[1] Earlier, Roussas (1969) had proven under stronger conditions a similar but weaker result for stationary Markov sequences. This was later improved in Roussas (1991a), though still under more restricted conditions (e.g. existence of densities) than Stute's (1986) in the i.i.d. case. Other estimates of conditional distribution functions in the context of Markov sequences have also been considered by Yakowitz (1979); see also Yakowitz (1985).

The list of authors proving convergence theorems for R_n is much bigger; we refer the reader to the bibliographic list of Collomb (1985), the books of Györfi et al. (1989), Härdle (1989), Roussas (1991b), Bosq (1998), and the article of Ango Nze and Doukhan (2002), for background and references to some of the most important and recent contributions in the area of nonparametric regression. Let us just mention that Stute (1986) showed that $R_n(u) \to R(u)$ a.s. for a.a. u in the i.i.d. case under what is the weakest condition known to this day, namely the existence of an absolute moment of order $r > 1$ and (essentially) $n^{1-1/r} h_n^m / \log n \to \infty$; and that Roussas (1990) and L.T. Tran (the latter cited by the former) proved the corresponding result for strictly stationary mixing processes, though under a more restrictive choice of the sequence $\{h_n\}$ and assuming, among other things, the existence of joint densities.

As suggested by the above sample of references, the work on nonparametric estimation in the context of discrete time processes concerns mainly strictly stationary and stationary Markov processes. The only exceptions seem to be the work of Collomb (1984) and its significant improvement and amplification by Györfi et al. (1989). The work of Collomb (1984) is, we think, quite important: He essentially realized that, apart from some form of ergodicity, in order to obtain convergence results for regression-type estimators such as (1.3) and (1.4) it is sufficient that there be (i) a regression function and (ii) a measure dominating the distributions of the U_i. This idea implies that consistency results require no stationarity assumption, and in fact that Markov and stationary processes can be both treated as special cases of what we have just called conditionally stationary processes.

In spite of its reworking and generalization by Györfi et al. (1989), Collomb's work has been somewhat forgotten. This is partly due to the shift of research into rates of convergence, central limit theorems, bandwidth selection methods and other topics involving more detailed

[1] Throughout, limits are taken as $n \to \infty$ unless stated otherwise.

information about convergence and which naturally demand stronger assumptions. However, since in actual practice the choice of parameters optimizing prediction is successfully achieved via rather empirical methods, and rates of convergence and similar results are often expressed in terms of practically incalculable items, it seems of great interest even today to weaken and clarify the conditions under which consistency alone is possible.

In Section 2 of this paper we obtain the strong consistency of the estimators (1.3) and (1.4) for conditionally stationary processes satisfying a strong mixing condition. Roughly speaking, Theorem 2.1 states that $\sup_{v\in\mathbb{R}} |F_n(v|u) - F(v|u)| \to 0$ a.s. under no assumptions on $(u, v) \to F(v|u)$, generalizing both Stute's and Roussas's results. In Theorem 2.2 we essentially prove $R_n(u) \to R(u)$ whenever $\sup_i E(|V_i|^r) \le M$ for some $M > 0$ and some $r > 1$, and in Theorem 2.3 a uniform version of this statement; these results generalize several of the available results in nonparametric regression, and partly those of Collomb (1984) and Györfi et al. (1989).

It is perhaps interesting to note that the consideration of conditionally stationary sequences (instead of stationary and Markov sequences separately, for instance) renders simple and transparent assumptions and relatively short proofs; moreover, once Bosq's inequality and Devroye's lemma (two of our main tools) are accepted, the proofs are practically self-contained.

The rest of the paper is devoted to more conceptual aspects of the prediction method based on a fixed number of past observations. In Section 3 we look a little further into the definition of conditionally stationary processes, and examine the question of whether nonparametric methods can be expected to behave in a nearly 'optimal way'. For the class of *approximately Markov processes* introduced there, this is shown to be true with large samples, and in a certain sense. The concepts introduced in Section 3 are illustrated by numerical results based on an ARMA$(1, 1)$ process in Section 4; these results answer in particular certain questions regarding the interplay between the sample size n and the number m of previous observations used to produce forecasts, and appear to be relevant in view of the increasing availability of large data sets in many scientific areas. Finally, Section 5 contains a statistical analysis of some wave data which clearly do not satisfy any form of stationarity; besides giving prediction results we informally check the assumption of conditional stationarity.

Before proceeding we need to clear one possible doubt: are there any non-trivial conditionally stationary sequences (i.e., non-stationary and non-Markov)? Although it does not seem so easy to come up with 'natural' examples of such processes, we do have a couple of simple ones (variations of which are possible); they serve to show in particular that our consistency results are not just a vague generalization of known ones—they may actually be applied to models not previously considered in the literature.[2]

Let $X^{(1)}, X^{(2)}, \ldots$ be a family of strictly stationary sequences with the same marginal distribution function H and such that the joint distribution function of $\left(X_k^{(i)}, X_{k+1}^{(i)}\right)$ is independent of i (such processes undoubtedly abound). Then $X^{(1)}, X^{(2)}, \ldots$ all have the same conditional distribution function $(u, v) \to F(v|u)$. Given a sequence $1 < k_1 < k_2 < \cdots$ of integers and a sequence $\xi_1, \xi_2, \ldots$ of independent, standard uniform random variables, we can generate a sequence X as follows (rather than the processes $X^{(1)}, X^{(2)}, \ldots$ themselves, what we use in this

[2]However, the establishment of conditions under which examples of conditionally stationary processes are mixing is presently beyond our scope.

construction is their family of joint distributions). Letting $H_{i,n}(\,\cdot\,|X_{n-1}^{(i)},\ldots,X_1^{(i)})$ stand for the conditional distribution function of $X_n^{(i)}$ given $X_{n-1}^{(i)},\ldots,X_1^{(i)}$, so that in particular $F(\,\cdot\,|u) = H_{i,2}(\,\cdot\,|u)$ a.a. u, we put $X_1 = H^{-1}(\xi_1)$ and $X_n = H_{1,n}^{-1}(\xi_n|X_{n-1},\ldots,X_1)$ for $1 < n \le k_1$; $X_n = H_{2,n-k_1+1}^{-1}(\xi_n|X_{n-1},\ldots,X_{k_1})$ for $k_1 < n \le k_2$; $X_n = H_{3,n-k_2+1}^{-1}(\xi_n|X_{n-1},\ldots,X_{k_2})$ for $k_2 < n \le k_3$; and so on. By construction, X has marginal distribution function H and conditional distribution function $(u,v) \to F(v|u)$; however, it is generally non stationary, since the segments $(X_1,\ldots,X_{k_1}),(X_{k_1+1},\ldots,X_{k_2}),\ldots$ may have a very different structure, and non Markov.

Next, suppose $(u,v) \to F(v|u)$ is defined for u and v on the same countable subset of $\mathbb{R}$, and that the associated Markov chain has a recurrent state u_0. For each $i \in \mathbb{N}$ let $X^{(i)}$ be a strictly stationary sequence with marginal distribution function $v \to F(v|u_0)$ and conditional distribution function $(u,v) \to F(v|u)$, and having u_0 also as a recurrent state. Writing $\tau_k = \min\{n > \tau_{k-1} : X_n = u_0\}$ for $k \in \mathbb{N}$ and $\tau_0 = 1$, and using the same notation as above, we set $X_1 = F^{-1}(\xi_1|u_0)$ and $X_n = F^{-1}(\xi_n|X_{n-1})$ for $1 < n \le \tau_1$; $X_n = H_{1,n-\tau_1+1}^{-1}(\xi_n|X_{n-1},\ldots,X_{\tau_1})$ for $\tau_1 < n \le \tau_2$; $X_{\tau_2+1} = F^{-1}(\xi_{\tau_2+1}|u_0)$ and $X_n = F^{-1}(\xi_n|X_{n-1})$ for $\tau_2+1 < n \le \tau_3$; $X_n = H_{3,n-\tau_3+1}^{-1}(\xi_n|X_{n-1},\ldots,X_{\tau_3})$ for $\tau_3 < n \le \tau_4$; and so on. Then the sequence X has the same conditional distribution but not necessarily the same marginal distribution, and it need not be stationary nor Markov.

2 Consistency

Throughout this section X is assumed *strongly mixing*: there exists a sequence $\alpha(1),\alpha(2),\ldots$ converging to zero such that

$$|P(A\cap B) - P(A)P(B)| \le \alpha(k) \tag{2.1}$$

for all $n \in \mathbb{N}$ and all sets $A \in \mathcal{F}_{1,n}$, $B \in \mathcal{F}_{n+k,\infty}$, where $\mathcal{F}_{1,n}$ and $\mathcal{F}_{n+k,\infty}$ are the sigma fields generated by $\{X_1,\ldots,X_n\}$ and $\{X_{n+k},X_{n+k+1},\ldots\}$, respectively. The numbers $\alpha(1),\alpha(2),\ldots$ are called mixing coefficients.

A basic tool for obtaining consistency results for strongly mixing sequences is the following 'exponential-type' inequality of Bosq (see p. 28 of Bosq (1998)):

Lemma 2.1. *Let $Y_1,Y_2,\ldots$ be a strongly mixing sequence with $E(Y_i) = 0$ and $|Y_i| \le b\ \forall i \in \mathbb{N}$, and mixing coefficients $\beta(1),\beta(2)\ldots$ Then for each $\varepsilon > 0$ and each integer $q \in [1, n/2]$,*

$$P\left(\left|\sum_{i=1}^{n} Y_i\right| > n\,\varepsilon\right) \le 4\exp\left\{-\frac{\varepsilon^2}{8\,\tau^2(q)}q\right\} + 22\left(1+\frac{4\,b}{\varepsilon}\right)^{1/2} q\,\beta\left(\left[\frac{n}{2q}\right]\right), \tag{2.2}$$

where

$$\tau^2(q) = \frac{2}{p^2}\sigma^2(q) + \frac{b\,\varepsilon}{2}, \tag{2.3}$$

with $p = \frac{n}{2q}$ and $\sigma^2(q) = \max_{j=0,1,\ldots,2q-1} E\Big[\Big(([jp]+1-jp)X_{[jp]+1} + X_{[jp]+2} + \cdots + X_{[(j+1)p]} + ((j+1)p - [j+1)p])\,X_{[(j+1)p)+1]}\Big)^2\Big]$.

The strong mixing condition is one of the weakest of the so-called 'weak dependence' conditions. Its importance in our context comes from the fact that it yields (2.2), which in principle may be replaced by a similar inequality; thus it should not be hard to adapt our proofs to other dependence contexts where exponential-type inequalities are available.

We shall also need a useful lemma of Devroye (1981):

Lemma 2.2. *Denote by $B_\varepsilon(u)$ the $m-$dimensional ball with radius ε centred at u. Let μ be a positive, sigma-finite measure and $\{\varepsilon_n\}$ a sequence of positive numbers such that $n\varepsilon_n^m \to \infty$. For all $c > 0$,*

$$n\mu\left(B_{c\,\varepsilon_n}(u)\right) \to \infty \quad as \quad for\ \mu - almost\ all\ u.$$

Our consistency results require three main assumptions:

A0 There exists a sequence $\{q_n\}$ of integers such that $q_n \in \left[1, \frac{n-m}{2}\right]$ and

$$q_n^{3/2}\alpha\left(\left[\frac{n}{2\,q_n}\right]\right) = O\left(\left(n(\log n)^{1/2}(\log\log n)^{1+\delta}\right)^{-1}\right) \tag{2.4}$$

for some $\delta > 0$.

A1 There exists a measure μ with the following property: For each $u \in \mathbb{R}^m$ there exist a neighbourhood $\mathcal{N}(u)$ of u and two strictly positive numbers C_1 and C_2, such that

$$C_1\mu(B) \le P\left(U_i \in B\right) \le C_2\mu(B)$$

for all i and all Borel sets $B \subset \mathcal{N}(u)$.

A2 There exists a non-decreasing function $G : \mathbb{R}_+ \to [0,1]$ with $G(\infty) = 1$ such that

$$\sup_i P\left(|V_i| > v\right) \le 1 - G(v) \tag{2.5}$$

for all sufficiently large v and $\int_0^\infty v^r dG(v) < \infty$ for some $r > 1$.

As we shall see, **A0** combined either with (2.7), (2.12) or (2.17) below provides weaker conditions for the consistency of estimators than those obtained by Györfi et al. (1989), which for instance do not apply to polynomially decreasing mixing coefficients; the improvement comes partly from using Bosq's inequality in place of Carbon's (1983).

Let us try to interpret these assumptions. **A1** is weaker than assumption (A.1) on p. 24 of Györfi et al. (1989), which is due to Collomb (1984). (Collomb (1984) essentially took μ as Lebesgue measure and required the second inequality of **A1** to hold for all $B \in \mathcal{B}^m$.) To see what it entails, observe that **A1** is equivalent to the requirement that there be a measure μ with the property that for each compact subset $\mathsf{K} \subset \mathbb{R}^m$ there exist constants $C_{1,\mathsf{K}}$ and $C_{2,\mathsf{K}}$ such that $C_{1,\mathsf{K}}\mu(B) \le P(U_i \in B) \le C_{2,\mathsf{K}}\mu(B)$ for every i and every Borel subset B of K. Thus **A1** is equivalent to the requirement that the restrictions of the probability measures $P(U_i \in \cdot\,)$ to any compact set K be absolutely continuous with respect to each other as well as to the restriction of μ to K, so that to each i there corresponds a density $f_{i,\mathsf{K}}$ such that $P(U_i \in B) = \int_B f_{i,\mathsf{K}} d\mu$

for every Borel subset B of K. If for K we take $Q_k := [-k,k]^m$ and put $g_{i,k} = 1_{Q_k} f_{i,Q_k}$, then $\int_B g_{i,k} d\mu = \int_B g_{i,k'} d\mu$ for $B \subset Q_k$ if $k < k'$ and $g_{i,k}$ increases a.e. with k, so it follows by the monotone convergence theorem that to each i there exists a density $g^{(i)} := \lim_{k\to\infty} g_{i,k}$ in $\mathbb{R}^m$ such that

$$P(U_i \in B) = \int_B g^{(i)} d\mu \quad \forall\, B \in \mathcal{B}^m. \tag{2.6}$$

Thus **A1** implies (2.6).

Assumptions **A0** and **A1** are the most fundamental; they are used everywhere. Together they imply that X is recurrent in the following sense: for $\mu-$a.a. u, the process visits every neighbourhood $\mathcal{N}(u)$ of u infinitely often. This follows from the fact that $\sum_{i=m}^{n-1} 1_{[U_i \in S(u,h_n)]} \sim \sum_{i=m}^{n-1} P\left(U_i \in S(u,h_n)\right)$ a.s., proved in Theorem 2.1 below, since if $\{h_n\}$ satisfies (2.7), hence a fortiori $n\, h_n^m \to \infty$, we have a.s., for large enough n,

$$\sum_{i=m}^{n-1} 1_{[U_i \in \mathcal{N}(u)]} \geq \sum_{i=m}^{n-1} 1_{[U_i \in S(u,h_n)]} \sim \sum_{i=m}^{n-1} P\left(U_i \in S(u,h_n)\right) \geq (n-m)\mu\left(S(u,h_n)\right) C_1 \to \infty$$

by Lemma 2.2.

Assumption **A2** is used to prove the consistency of empirical regression functions. It is equivalent to the requirement that $\sup_i E\left(|V_i|^r\right) \leq M$ for some $M > 0$ and some $r > 1$. Indeed, the latter condition implies $\sup_i v^{r-\delta-1} P\left(|V_i| > v\right) \leq M\, v^{-(\delta+1)}$ for all $\delta, v > 0$, and therefore that the function G defined by $G(v) = \inf_i P\left(|V_i| \leq v\right)$ $(v \geq 0)$ is such that $G(\infty) = 1$ and $\int_0^\infty v^{r-\delta} dG(v) < \infty$ for all $\delta \in (0, r)$, and hence satisfies **A2**. Conversely, (2.5) and $\int_0^\infty v^r dG(v) < \infty$ imply $\sup_i E\left(|V_i|^r\right) \leq M$ for some $M \geq 0$.

Theorem 2.1. *Suppose* (i) $\{q_n\}$ *is a sequence satisfying* **A0**, (ii) **A1** *holds. If* $\{h_n\}$ *is a sequence of positive numbers converging to zero in such a way that*

$$\frac{q_n\, h_n^m}{\log n} \to \infty, \tag{2.7}$$

then for $\mu-$*a.a.* u

$$\sup_{v\in\mathbb{R}} |F_n(v|u) - F(v|u)| \to 0$$

with probability 1.

Proof. We shall first prove that, for fixed v, $|F_n(v|u) - F(v|u)| \to 0$ a.s. for $\mu-$a.a. u, and later show this implies the conclusion of the theorem.

Put

$$\begin{aligned}
\Delta_{1,n} &= \sum_{i=m}^{n-1} \frac{\left\{1_{[V_i \leq v, U_i \in S(u,h_n)]} - P(V_i \leq v, U_i \in S(u,h_n))\right\}}{\sum_{j=m}^{n-1} P(U_j \in S(u,h_n))}, \\
\Delta_{2,n} &= \sum_{i=m}^{n-1} \frac{\left\{P(V_i \leq v | U_i \in S(u,h_n)) - F(v|u)\right\} P(U_i \in S(u,h_n))}{\sum_{j=m}^{n-1} P(U_j \in S(u,h_n))}, \\
\Delta_{3,n} &= \frac{\sum_{i=m}^{n-1} 1_{[U_i \in S(u,h_n)]}}{\sum_{i=m}^{n-1} P(U_i \in S(u,h_n))} - 1 = \sum_{i=m}^{n-1} \frac{\left\{1_{[U_i \in S(u,h_n)]} - P(U_i \in S(u,h_n))\right\}}{\sum_{j=m}^{n-1} P(U_j \in S(u,h_n))}.
\end{aligned}$$

Then

$$|F_n(v|u) - F(v|u)| \le |\Delta_{1,n}| + |\Delta_{2,n}| + \left|\frac{\Delta_{3,n}}{1+\Delta_{3,n}}\right| |\Delta_{1,n} + \Delta_{2,n} + F(v|u)|\,,$$

and we need to prove the convergence of $\Delta_{1,n}$, $\Delta_{2,n}$ and $\Delta_{3,n}$ to zero.

To prove $\lim_{n\to\infty} |\Delta_{2,n}| = 0$, we note that the existence of a common conditional distribution function and assumption **A1** give

$$\begin{aligned}\left|\frac{P(V_i \le v, U_i \in S(u,h)))}{P(U_i \in S(u,h)))} - F(v|u)\right| &\le \frac{1}{P(U_i \in S(u,h))} \int\limits_{S(u,h)} \left|F(v|u') - F(v|u)\right| dP(U_i \le u') \\ &\le \frac{C_2/C_1}{\mu(S(u,h))} \int\limits_{S(u,h)} \left|F(v|u') - F(v|u)\right| \mu(du'), \qquad (2.8)\end{aligned}$$

and by Lebesgue's density theorem this tends to zero as $h \downarrow 0$ for $\mu-$a.a. u.

Set $n' := n - m$, $\Pi_{n'} := \frac{1}{n'}\sum_{i=1}^{n'} P(U_{i+m-1} \in S(u,h_n))$, and define

$$Y_{i-m+1} := \Pi_{n'}^{-1}\left(1_{[V_i \le v, U_i \in S(u,h_n)]} - P(V_i \le v, U_i \in S(u,h_n))\right), \quad i \ge m.$$

Then $Y_1, Y_2, \ldots$ is a strongly mixing sequence with mixing coefficients given by $\beta(k) = \alpha(k-m)$ for $k > m$ and $\beta(k) = 1$ for $1 \le k \le m$, and satisfying $|Y_{i-m+1}| \le \Pi_{n'}^{-1}$. Thus Bosq's inequality (2.2) applies to give, for sufficiently large n,

$$\begin{aligned}P\left(|\Delta_{1,n}| > \varepsilon\right) &= P\left(\left|\sum\nolimits_{i=1}^{n'} Y_i\right| > n'\,\varepsilon\right) \\ &\le 4\exp\left\{-\frac{\varepsilon^2}{8\,\tau^2(q_{n'})} q_{n'}\right\} + 22\left(1 + \frac{4}{\varepsilon\,\Pi_{n'}}\right)^{1/2} q_{n'}\,\beta\left(\left[\frac{n'}{2q_{n'}}\right]\right) \qquad (2.9)\end{aligned}$$

for every $\varepsilon > 0$ and integer $q_{n'} \in [1, n'/2]$, where $\tau^2(q_{n'})$ is defined by (2.3) with $p_{n'} := \frac{n'}{2q_{n'}}$ and $q_{n'}$ in place of p and q, respectively.

Furthermore, for large enough n we have by assumption **A1**

$$\left|\,E[Y_{i-m+1}Y_{j-m+1}]\,\right| \le \frac{1}{\Pi_{n'}^2} P(U_i \in S(u,h_n)) = \frac{1}{\Pi_{n'}} \frac{P(U_i \in S(u,h_n))}{\frac{1}{n'}\sum_{r=1}^{n'} P(U_{r+m-1} \in S(u,h_n))} \le \frac{1}{\Pi_{n'}} c_1,$$

for all $i, j \ge m$ and a constant $c_1 > 0$. This inequality yields

$$\sigma^2(q_{n'}) \le 2\,c_1 \frac{\left(2p_{n'} + 1 + \frac{p_{n'}(p_{n'}-1)}{2}\right)}{\Pi_{n'}} \le 2\,c_1 \frac{(p_{n'}+1)^2}{\Pi_{n'}},$$

and, since $1 + \frac{1}{p_{n'}} \le 2$,

$$\tau^2(q_{n'}) = \frac{2}{p_{n'}^2}\sigma^2(q_{n'}) + \frac{\varepsilon}{2\,\Pi_{n'}} \le \frac{1}{\Pi_{n'}}\left(16\,c_1 + \frac{\varepsilon}{2}\right) \le \frac{c_2}{\Pi_{n'}}\,(1+\varepsilon)$$

for some $c_2 > 0$. Using this in (2.9) we conclude that for sufficiently large n there is a constant $c > 0$ such that

$$P\left(|\Delta_{1,n}| > \varepsilon\right) \leq 4 \exp\left\{-c \frac{\varepsilon^2}{1+\varepsilon} q_{n'} \Pi_{n'}\right\} + 22 \left(1 + \frac{4}{\varepsilon\, \Pi_{n'}}\right)^{1/2} q_{n'}\, \beta\left(\left[\frac{n'}{2q_{n'}}\right]\right) \tag{2.10}$$

for all $\varepsilon > 0$.

We shall now show that (2.4) and (2.7) imply the summability of both terms on the right-hand side of (2.10), and hence that $\lim_{n\to\infty} |\Delta_{1,n}| = 0$ a.s.

For the first term to be summable it is sufficient that $q_{n'} \Pi_{n'} / \log n' \to \infty$. By assumption **A1**, $\Pi_{n'} \geq C_1 \mu(S(u, h_n))$ for large enough n, so by Devroye's lemma it is sufficient to require $q_{n'}\, h_n^m / \log n' \to \infty$, which is (2.7).

Since either $1/\Pi_{n'} \to \infty$ or $1/\Pi_{n'} = O(1)$, we have $\left(1 + \frac{4}{\varepsilon\, \Pi_{n'}}\right)^{1/2} = O\left(\Pi_{n'}^{-1/2}\right)$, so the summability of the second term on the right of (2.10) is equivalent to

$$\sum_{n'} \frac{1}{(\Pi_{n'} q_{n'} / \log n')^{1/2}} \frac{q_{n'}^{3/2}}{(\log n')^{1/2}}\, \beta\left(\left[\frac{n'}{2q_{n'}}\right]\right) < \infty,$$

for which (2.4) is sufficient.

The proof of $\lim_{n\to\infty} |\Delta_{3,n}| = 0$ a.s. is practically the same.

Now let $\varepsilon > 0$. By a classical argument (e.g. Tucker (1967), pp. 127-128), we know that for a given $u \in \mathbb{R}^m$ there exist points $-\infty \leq v_0 < v_1 < \cdots < v_{k(\varepsilon)} \leq \infty$ such that $F(v_i|u) - F(v_{i-1}|u) \leq \varepsilon$, $i = 1, \ldots, k(\varepsilon)$. If $v \in [v_{i-1}, v_i]$, then

$$F_n(v_{i-1}|u) - F(v_{i-1}|u) - \varepsilon \leq F_n(v|u) - F(v|u) \leq F_n(v_i|u) - F(v_i|u) + \varepsilon,$$

so

$$\sup_v |F_n(v|u) - F(v|u)| \leq \max_{i=1,\ldots,k(\varepsilon)} |F_n(v_i|u) - F(v_i|u)| + \varepsilon.$$

If $|F_n(v_i|u) - F(v_i|u)| \not\to 0$ a.s. for some i, then $u \in S_\varepsilon$ for some Borel set with $\mu(S_\varepsilon) = 0$. Otherwise,

$$\limsup_{n\to\infty} \sup_v |F_n(v|u)_\omega - F(v|u)| \leq \varepsilon$$

for ω on a set $\Omega_\varepsilon \subset \mathcal{F}$ with $P(\Omega_\varepsilon) = 1$.

Finally, set $S := \bigcup_{\varepsilon\in\mathbb{Q}\,:\,\varepsilon>0} S_\varepsilon$ and $\Omega_0 := \bigcap_{\varepsilon\in\mathbb{Q}\,:\,\varepsilon>0} \Omega_\varepsilon$ with S_ε and Ω_ε defined as above. Then $\mu(S) = 0$, $P(\Omega_0) = 1$, and for $u \in S^C$, $\omega \in \Omega_0$,

$$\lim_{n\to\infty} \sup_v |F_n(v|u)_\omega - F(v|u)| = 0,$$

which proves the result. □

If the pairs $(U_1, V_1), (U_2, V_2), \ldots$ are independent, the assumptions of the theorem are satisfied with $q_n = [n/2]$ and (2.7) becomes $n\, h_n^m / \log n \to \infty$, essentially Stute's (1986) condition.

If the mixing coefficients decrease exponentially, i.e., if $\alpha(k) = O(e^{-\theta\, k})$ for some $\theta > 0$, then $q_n = [c\, n / \log n]$ with $c \geq \frac{5}{2\theta}$ satisfies (2.4), and the conclusion of the theorem holds with

any $\{h_n\}$ converging to zero in such a way that $n\,h_n^m/(\log n)^2 \to \infty$. This condition coincides with Collomb's (1984) in the less general case of $\varphi-$mixing sequences, and is only slightly more stringent than Stute's.

Suppose the mixing coefficients decrease polynomially: $\alpha(k) = O(k^{-\gamma})$ for some $\gamma > 0$. Then the conclusion of the theorem is true if

$$\frac{n^{2(\gamma-1)/(2\gamma+3)}}{\left\{(\log n)^{1/2}(\log\log n)^{1+\delta}\right\}^{2/(2\gamma+3)}} h_n^m \to \infty,$$

which requires $\gamma > 1$ and naturally is the more stringent the slower the decay of α is.

Similar statements apply to Theorem 2.2 and Theorem 2.3.

Another consequence of the theorem is the following: If $F(\,\cdot\,|u)$ is strictly increasing at $x_\epsilon(u) := F^{-1}(\epsilon|u) \equiv \min\{x : F(x|u) \geq \epsilon\}$, the conditional quantile of probability $\epsilon \in (0,1)$, or equivalently if $F^{-1}(\,\cdot\,|u)$ is continuous at ϵ, then

$$x_\epsilon^{(n)}(u) := F_n^{-1}(\epsilon|u) \to x_\epsilon(u)$$

with probability 1, i.e., the $\epsilon-$*conditional sample quantile* $x_\epsilon^{(n)}(u)$ (the quantile of probability ϵ of the empirical conditional distribution function) converges a.s. to $x_\epsilon(u)$. Thus, for example, if $F(\,\cdot\,|u)$ is strictly increasing and continuous at $x_{\epsilon/2}(u)$ and $x_{1-\epsilon/2}(u)$, then

$$P\Big(X_{n+1} \in \big(x_{\epsilon/2}^{(n)}(u), x_{1-\varepsilon/2}^{(n)}(u)\big)\Big|(X_{n-1},\ldots,X_{n-m}) = u\Big) \to 1-\varepsilon, \tag{2.11}$$

a statement which may justify the use of the $(1-\varepsilon)\%$ *predictive interval* $\big(x_{\epsilon/2}^{(n)}(u), x_{1-\varepsilon/2}^{(n)}(u)\big)$ in practice.

Let us remark that Theorem 2.1 is a statement about a.s. convergence, while the consistency of non-parametric estimators is sometimes proved in the sense of complete convergence. Though not stated, the convergence in our two other results is in the complete sense.

Theorem 2.2. *Suppose* (i) $\{q_n\}$ *is a sequence satisfying* **A0**, (ii) **A1** *and* **A2** *hold. If* $\{h_n\}$ *is a sequence of positive numbers converging to zero in such a way that*

$$\frac{q_n\,h_n^m}{n^{1/r}\log n} \to \infty, \tag{2.12}$$

then for $\mu-$*a.a.* u

$$R_n(u) \to R(u)$$

with probability 1.

Proof. Let us observe in the first place that assumption **A2** implies the existence of $E\,[|V_i|]$, and hence of $E\left[|V_i|\big|U_i = u\right]$, for all i.

Taking up the notation in the proof of Theorem 2.1, we define

$$\begin{aligned}
\Delta_{1,n}^* &= \sum_{i=m}^{n-1} \frac{\left\{V_i 1_{[U_i\in S(u,h_n)]} - E[V_i 1_{[U_i\in S(u,h_n)]}\right\}}{n'\Pi_{n'}}, \\
\Delta_{2,n}^* &= \sum_{i=m}^{n-1} \frac{\left\{E[V_i|U_i \in S(u,h_n)] - R(u)\right\} P(U_i \in S(u,h_n))}{n'\Pi_{n'}},
\end{aligned}$$

and $\Delta^*_{3,n} = \Delta_{3,n}$. Then

$$|R_n(u) - R(u)| \le |\Delta^*_{1,n}| + |\Delta^*_{2,n}| + \left|\frac{\Delta_{3,n}}{1+\Delta_{3,n}}\right| \left|\Delta^*_{1,n} + \Delta^*_{2,n} + R(u)\right|,$$

and we have to show that $\left|\Delta^*_{1,n}\right| \to 0$ a.s. and $\left|\Delta^*_{2,n}\right| \to 0$ for $\mu-$a.a. u.

Take $\varepsilon > 0$ arbitrary, and choose T so large that $\int_T^\infty 1 - F(v|u) + F(-v|u)\, dv < \varepsilon' := \varepsilon/(1 + C_2/C_1)$. Then

$$\begin{aligned}\left|E[V_i|U_i \in S(u,h_n)] - R(u)\right| &\le \int_{-T}^{T} \left|F(v|u) - P\left(V_i \le v | U_i \in S(u,h_n)\right)\right| dv + \varepsilon' + \\ &\quad \int_T^\infty P\left(|V_i| > v | U_i \in S(u,h_n)\right) dv.\end{aligned}$$

By (2.8), for $\mu-$a.a. u the first term on the right hand side of this inequality tends to zero as $n \to \infty$ uniformly in i. As to the last term, note that

$$\begin{aligned}\int_T^\infty P\left(|V_i| > v | U_i \in S(u,h_n)\right) dv &= \int_T^\infty \int_{S(u,h_n)} 1 - F(v|u') + F(-v|u') \frac{dP(U_i \le u')}{P\left(U_i \in S(u,h_n)\right)} dv \\ &\le \frac{C_2}{C_1} \int_{S(u,h_n)} \int_T^\infty 1 - F(v|u') + F(-v|u')\, dv \frac{\mu(du')}{\mu\left(S(u,h_n)\right)},\end{aligned}$$

hence

$$\limsup_{n\to\infty} \max_{m \le i \le n-1} \int_T^\infty P\left(|V_i| > v | U_i \in S(u,h_n)\right) dv < \varepsilon' C_2/C_1 \quad (\mu - \text{a.a. } u). \tag{2.13}$$

Thus

$$\limsup_{n\to\infty} \max_{m \le i \le n-1} \left|E[V_i|U_i \in S(u,h_n)] - R(u)\right| < \varepsilon \quad (\mu - \text{a.a. } u), \tag{2.14}$$

which proves $\left|\Delta^*_{2,n}\right| \to 0$ for $\mu-$a.a. u.

Now let $\{T_n\}$ be a sequence of positive numbers increasing to infinity, and define

$$Y^*_{i-m+1} := \Pi_{n'}^{-1} \left(V_i 1_{[|V_i| \le T_n, U_i \in S(u,h_n)]} - E[V_i 1_{[|V_i| \le T_n, U_i \in S(u,h_n)]}]\right), \quad i \ge m.$$

Then

$$\begin{aligned}\left|\Delta^*_{1,n}\right| &\le \left|\frac{1}{n}\sum_{i=m}^{n-1} Y^*_{i-m+1}\right| + \frac{1}{n}\sum_{i=m}^{n-1} \frac{E\left[|V_i| 1_{[|V_i| > T_n, U_i \in S(u,h_n)]}\right]}{\Pi_{n'}} + \frac{1}{n}\sum_{i=m}^{n-1} \frac{|V_i| 1_{[|V_i| > T_n, U_i \in S(u,h_n)]}}{\Pi_{n'}} \\ &=: \delta_{1,n} + \delta_{2,n} + \delta_{3,n},\end{aligned}$$

and we want to prove the convergence of $\delta_{1,n}$, $\delta_{2,n}$ and $\delta_{3,n}$ to zero.

As in the proof of Theorem 2.1, we see that $Y^*_1, Y^*_2, \ldots$ is a strongly mixing sequence with mixing coefficients given by $\beta(k) = \alpha(k-m)$ for $k > m$ and $\beta(k) = 1$ for $1 \le k \le m$, and

satisfying $|Y^*_{i-m+1}| \le 2\,T_n/\Pi_{n'}$. Moreover, by **A1** we have, for n large enough,

$$\begin{aligned}
\big|\, E[Y^*_{i-m+1}Y^*_{j-m+1}]\,\big| &\le \frac{T_n}{\Pi^2_{n'}} E\,\big|\, V_i 1_{[|V_i|>T_n, U_i\in S(u,h_n)]} - E\left(V_i 1_{[|V_i|>T_n, U_i\in S(u,h_n)]}\right)\big| \\
&= \frac{T_n}{\Pi^2_{n'}} \int_{S(u,h_n)} E\big[\,|V_i|\big|U_i = u'\big]\, dP(U_i \le u') + \frac{T_n}{\Pi^2_{n'}} E\left(|V_i| 1_{[U_i\in S(u,h_n)]}\right) \\
&\le \frac{c_0 T_n}{\Pi_{n'}} \left\{ \frac{1}{\mu\left(S(u,h_n)\right)} \int_{S(u,h_n)} E\big[\,|V_i|\big|U_i = u'\big]\,\mu(du') + \frac{E\left(|V_i| 1_{[U_i\in S(u,h_n)]}\right)}{\Pi_{n'}} \right\}
\end{aligned}$$

for some $c_0 > 0$. Because the integrand in the rightmost term of these inequalities is independent of i, we may conclude from Lebesgue's density theorem and from equation (2.14) that $\max_{m\le i\le n-1} \big|\, E[Y^*_{i-m+1}Y^*_{j-m+1}]\,\big| \le c_1 T_n/\Pi_{n'}$ for some $c_1 > 0$ ($\mu-$a.a. u), which in turn yields, as in the proof of Theorem 2.1, $\tau^2(q_{n'}) \le c_2\, T_n/\Pi_{n'}$ for some $c_2 > 0$ ($\mu-$a.a. u). Applying Bosq's inequality we thus find that for $\mu-$a.a. u and sufficiently large n there is a constant $c > 0$ such that

$$P\left(\delta_{1,n} > \varepsilon\right) \le 4\exp\left\{-c\frac{\varepsilon^2}{1+\varepsilon}\frac{q_{n'}\Pi_{n'}}{T_n}\right\} + 22\left(1 + \frac{8\,T_n}{\varepsilon\,\Pi_{n'}}\right)^{1/2} q_{n'}\,\beta\left(\left[\frac{n'}{2q_{n'}}\right]\right)$$

for all $\varepsilon > 0$. If we now choose $T_n = n^{1/r}$, where $r > 1$ is such that $\int_0^\infty v^r dG(v) < \infty$, then it follows essentially as in the proof of Theorem 2.1 that (2.4) and (2.12) imply the summability of both terms on the right-hand side of this last inequality. This proves $\lim_{n\to\infty}\delta_{1,n} = 0$ a.s.

Next, note that if $\varepsilon > 0$ is given and T is chosen sufficiently large then, by (2.13),

$$\begin{aligned}
\delta_{2,n} &\le \frac{C_2}{C_1}\frac{1}{n}\sum_{i=m}^{n-1} \frac{E\left[|V_i| 1_{[|V_i|>T_n, U_i\in S(u,h_n)]}\right]}{P\left(U_i\in S(u,h_n)\right)} = \frac{C_2}{C_1}\frac{1}{n}\sum_{i=m}^{n-1}\int_{T_n}^\infty P\left(|V_i| > v | U_i\in S(u,h_n)\right) dv \\
&\le \frac{C_2}{C_1}\max_{m\le i\le n-1}\int_T^\infty P\left(|V_i| > v | U_i\in S(u,h_n)\right) dv < \varepsilon
\end{aligned}$$

for all sufficiently large n. This proves $\lim_{n\to\infty}\delta_{2,n} = 0$ for $\mu-$a.a. u.

Finally, because $\lim_{n\to\infty}\Delta_{3,n} = 0$ (Theorem 2.1), we can show that $\lim_{n\to\infty}\delta_{3,n} = 0$ a.s. by showing instead that

$$\sum_{i=m}^{n-1} \frac{|V_i| 1_{[|V_i|>T_n, U_i\in S(u,h_n)]}}{\sum_{j=m}^{n-1} 1_{[U_j\in S(u,h_n)]}} \le \sum_{i=m}^{n-1} \frac{|V_i| 1_{[|V_i|>T_i]} 1_{[U_i\in S(u,h_n)]}}{\sum_{j=m}^{n-1} 1_{[U_j\in S(u,h_n)]}} \to 0 \quad \text{(a.s.)}. \tag{2.15}$$

By Lemma 6.1.1 of Ash and Doléans-Dade (2000), we know that if $\{x_i\}$ is a sequence converging to zero and $\{a_{ni}\}$ a double sequence such that $\lim_{n\to\infty} a_{ni} = 0$ for each i and $\sum_{j=1}^\infty a_{nj} = 1$, then $\lim_{n\to\infty}\sum_{i=1}^\infty x_i a_{ni} = 0$. Let $a_{ni} = 1_{[U_i\in S(u,h_n)]}/\sum_{j=m}^{n-1} 1_{[U_j\in S(u,h_n)]}$ if $m \le i \le n-1$ and $a_{ni} = 0$ otherwise, and $x_i = |V_i| 1_{[|V_i|>T_i]}$. Observe that by **A1**, Devroye's lemma and (2.12) we have with probability 1

$$\frac{1_{[U_i\in S(u,h_n)]}}{\sum_{j=m}^{n-1} 1_{[U_i\in S(u,h_n)]}} \sim \frac{1_{[U_i\in S(u,h_n)]}}{\sum_{j=m}^{n-1} P\left(U_i\in S(u,h_n)\right)} \le \frac{C_2}{C_1}\frac{1}{n\,\mu\left(S(u,h_n)\right)} \to 0,$$

so $\lim_{n\to\infty} a_{ni} = 0$ a.s., and we thus see that (2.15) will follow once we prove that $\{x_i\}$ converges to zero a.s., or equivalently that $\lim_{i\to\infty} 1_{[|V_i|>T_i]} = 0$ a.s., or, still equivalently, that $P(|V_i| > T_i \text{ i.o.}) = 0$. But, by **A2**,

$$\sum_{i=1}^{\infty} P(|V_i| > T_i) \leq \sum_{i=1}^{\infty} 1 - G(T_i),$$

and (e.g. Knopp (1928), p. 294) with $T_n = n^{1/r}$ this last series is convergent if and only if

$$\int_0^{\infty} v^{r-1} \left[1 - G(v)\right] dv < \infty,$$

which in turn holds if and only if G has a finite moment of order r. □

Collomb (1984) and Györfi et al. (1989) proved a stronger type of statement, namely

$$\sup_{u\in\mathsf{K}} |R_n(u) - R(u)| \to 0 \quad \text{a.s.} \tag{2.16}$$

where $\mathsf{K} \subset \mathbb{R}^m$ is compact, assuming in particular the continuity of R and the existence of strictly positive Lebesgue densities for $P(U_i \in \cdot\,)$. Actually, the result they proved was for a 'smoothed' version of R_n in which the factors $1_{[U_i\in S(u,h_n)]}$ are replaced by $K\left((u - U_i)/h_n\right)$, K being a kernel satisfying a Lipschitz condition. Our version of (2.16) contained in Theorem 2.3 is again for the 'unsmoothed' version of the regression estimator (the proof of **(d)** below would be simpler in a smoothed version). Before stating it, we need to introduce further assumptions. In the following, λ denotes Lebesgue measure on $\mathbb{R}^m$.

A1* There exists a compact set $\mathsf{K} \subset \mathbb{R}^m$ and two strictly positive numbers $C_{1,\mathsf{K}}$ and $C_{2,\mathsf{K}}$ such that

$$C_{1,\mathsf{K}}\lambda(B) \leq P\left(U_i \in B\right) \leq C_{2,\mathsf{K}}\lambda(B)$$

for all i and all Borel sets $B \subset \mathsf{K}$.

A3 We have $\sup_{u\in\mathsf{K}} E\left[|V_i|^r \middle| U_i = u\right] = \sup_{u\in\mathsf{K}} \int |v|^r dF(v|u) < \infty$ for some $r > 1$.

A4 For every $T > 0$ and every $\varepsilon > 0$, there is $\delta > 0$ such that

$$\int_{-T}^{T} \left|F(v|u) - F(v|u')\right| dv < \varepsilon$$

for all u', u in K such that $|u - u|' < \delta$.

Remarks

(i) As pointed out by Bosq (1998), p. 72, one can not in general hope to obtain $R_n \to R$ uniformly over the whole of $\mathbb{R}^m$.

(ii) If the variables U_i are concentrated on the same lattice then (2.16) follows of course from Theorem 2.2.

(iii)**A4** holds under the condition that for each v outside a fixed set of Lebesgue measure zero the mapping $u \to F(v|u)$ is continuous on K.

(iv) For smoothed versions of R_n, results like (2.16) can apply only to continuous R. Our assumptions do not seem to imply the continuity of R on K. But they do if instead of **A4** (or instead of the sufficient condition stated in (iii)) we assume that $F(\cdot\,|u_n)$ converges weakly to $F(\cdot\,|u)$ whenever $u_n \to u$ in K. [Proof: If $\xi_{u_1}, \xi_{u_2}, \ldots$ is a sequence with distribution functions $F(\cdot\,|u_1), F(\cdot\,|u_2), \ldots$, then, by **A3**, $\sup_n E\left[|\xi_{u_n}|^r\right] \le M$ for some $M \ge 0$, so $\{\xi_{u_n}\}$ is uniformly integrable and therefore $E\left[\xi_{u_n}\right] \to E\left[\xi_u\right]$, where ξ_u has distribution function $F(\cdot\,|u)$, which is the same as $R(u_n) \to R(u)$.]

Theorem 2.3. *Suppose* (*i*) $\{q_n\}$ *is a sequence satisfying* **A0**, (*ii*) **A1*** *holds for some compact* $\mathsf{K} \subset \mathbb{R}$, (*iii*) **A2** *and* **A3** *hold with some* $r > 1$, (*iv*) **A4** *holds. If* $\{h_n\}$ *is a sequence of positive numbers converging to zero in such a way that* [3]

$$\frac{q_n\, h_n^{m(1+2\gamma)}}{n^{1/r}\log n} \to \infty \tag{2.17}$$

for some $\gamma > 1$, *then*

$$\sup_{u\in\mathsf{K}} |R_n(u) - R(u)| \to 0$$

with probability 1.

Proof. As we shall make use of results already proved, it will be convenient to make explicit the dependence of the remainder terms in the proof of Theorem 2.2 on the variable u by denoting them by $\Delta^*_{1,n}(u)$, $\Delta^*_{2,n}(u)$, $\Delta^*_{3,n}(u)$, $\delta_{1,n}(u)$, $\delta_{2,n}(u)$ and $\delta_{3,n}(u)$.

We divide the proof in several steps.

(a) *Given* $\varepsilon > 0$, *we can choose* $T > 0$ *such that*

$$\sup_{u\in\mathsf{K}} \int_T^\infty 1 - F(v|u) + F(-v|u)dv < \varepsilon.$$

Writing ξ_u for a random variable with distribution function $F(\cdot\,|u)$, we see that the integral in question is

$$\int_{\{|\xi_u|>T\}} |\xi_u|\, dP \le \frac{\sup_{u\in\mathsf{K}} E\left[|\xi_u|^\delta\right]}{T^{\delta-1}} \le \frac{M}{T^{\delta-1}}$$

for some $M > 0$, by **A3**.

(b) $\sup_{u\in\mathsf{K}} \left|\Delta^*_{2,n}(u)\right| \to 0$.

We have

$$\sup_{u\in\mathsf{K}} \left|\Delta^*_{2,n}(u)\right| \le \frac{C_{2,\mathsf{K}}}{C_{1,\mathsf{K}}} \sup_{u\in\mathsf{K}} \max_{m\le i\le n-1} \left|E\left[V_i | U_i \in S(u,h_n)\right] - R(u)\right|, \tag{2.18}$$

[3] It will be clear from the proof that Theorem 2.3 effectively holds with $n^{1/r}\log n/\left(q_n h_n^{m(1+2\gamma)}\right) = O(1)$ in place of the slightly more stringent condition (2.17), but the latter is more in line with (2.7) and (2.12).

and we shall show this last term goes to zero. By **(a)**, given $\varepsilon > 0$ we can choose $T > 0$ such that $\sup_{u\in\mathsf{K}} \int_T^\infty 1 - F(v|u) + F(-v|u)dv < \varepsilon := \varepsilon/\left(1 + C_{2,\mathsf{K}}/C_{1,\mathsf{K}}\right)$, hence, as in the argument around (2.13), the right hand of (2.18) is bounded above by a constant times

$$\sup_{u\in\mathsf{K}} \max_{m\le i\le n-1} \int_{-T}^{T} \left|F(v|u) - P\left(V_i \le v | U_i \in S(u,h_n)\right)\right| dv + \varepsilon.$$

But the inequality (2.8) and **A4** imply, by the bounded convergence theorem, that this last integral can be made arbitrarily small by taking n large enough, and this proves our statement.

(c) $\sup_{u\in\mathsf{K}} \delta_{1,n}(u) \to 0$ *a.s.*

Following Collomb (1984), p. 455, we let, for each $n \in \mathbb{N}$, $S_1, \ldots, S_{l_n}$ be a cover of K by l_n squares centred at $u_1, \ldots, u_{l_n} \in \mathsf{K}$ with edges (parallel to the coordinate axes) of length at most h_n^γ, where $\gamma > 1$. Moreover, since K is bounded we can choose this cover in such a way that $l_n \le L_\mathsf{K} h_n^{-\gamma m}$ for some $L_\mathsf{K} \ge 0$. Put

$$\tilde{\delta}_{1,n}(u) = \delta_{1,n}(u) - \delta_{1,n}(u_k), \quad u \in S_k \cap \mathsf{K}.$$

Then

$$\delta_{1,n}(u) \le \delta_{1,n}(u_k) + \left|\tilde{\delta}_{1,n}(u)\right|, \quad u \in S_k \cap \mathsf{K},$$

and **(c)** will be proved once we show

$$\max_{1\le k\le l_n} \delta_{1,n}(u_k) \to 0 \quad \text{and} \quad \sup_{u\in\mathsf{K}} \left|\tilde{\delta}_{1,n}(u)\right| \to 0 \quad \text{a.s.} \tag{2.19}$$

Let us first note that the upper bound for $P(\delta_{1,n} > \varepsilon)$ obtained in the proof of Theorem 2.2 applies to $\delta_{1,n}(u)$ as well, and uniformly in K; specifically, for sufficiently large n there exist $c_1, c_2 > 0$ such that

$$\sup_{u\in\mathsf{K}} P(\delta_{1,n}(u) > \varepsilon) \le 4\exp\left\{-c_1 \frac{\varepsilon^2}{1+\varepsilon} \frac{q_{n'} h_n^m}{T_n}\right\} + 22\left(1 + \frac{c_2 T_n}{\varepsilon\, h_n^m}\right)^{1/2} q_{n'}\, \beta\left(\left[\frac{n'}{2q_{n'}}\right]\right) \tag{2.20}$$

for all $\varepsilon > 0$. (Recall $T_n = n^{1/r}$.)

Since $P\left(\max_{1\le k\le l_n} \delta_{1,n}(u_k)\right) \le L_\mathsf{K} h_n^{-\gamma m} \sup_{u\in\mathsf{K}} P\left(\delta_{1,n}(u)\right)$, in order to secure the first statement in (2.19) it remains to show that both terms obtained by multiplying the right hand side of (2.20) by $h_n^{-\gamma m}$ are summable. As to the first term, it is sufficient to require $q_{n'} h_n^m / (T_n \log n) \to \infty$ and $\liminf_n \log h_n / \log n > -\infty$, both of which are easily seen to be implied by (2.17). By **A0** the summability of the second requires only $T_n \log n / \left(q_n h_n^{m(1+2\gamma)}\right) = O(1)$, and this follows again by (2.17).

Next, writing $\tilde{S}_1(u,u_k,h_n) := S(u,h_n) \cap S(u_k,h_n)^C$, $\tilde{S}_2(u,u_k,h_n) := S(u,h_n)^C \cap S(u_k,h_n)$, and noting

$$1_{[U_i\in S(u,h_n)]} - 1_{[U_i\in S(u_k,h_n)]} = 1_{[U_i\in \tilde{S}_1(u,u_k,h_n)]} - 1_{[U_i\in \tilde{S}_2(u,u_k,h_n)]},$$

we see that

$$\begin{aligned}\left|\tilde{\delta}_{1,n}(u)\right| \;\leq\; & \frac{C_{1,\mathsf{K}}^{-1}}{nh_n^m}\sum_{i=m}^{n-1}\left|V_i 1_{[|V_i|\leq T_n]}1_{[U_i\in\tilde{S}_1(u,u_k,h_n)]}\right| + \frac{C_{1,\mathsf{K}}^{-1}}{nh_n^m}\sum_{i=m}^{n-1}\left|V_i 1_{[|V_i|\leq T_n]}1_{[U_i\in\tilde{S}_2(u,u_k,h_n)]}\right| + \\ & \frac{C_{1,\mathsf{K}}^{-1}}{nh_n^m}\sum_{i=m}^{n-1}E\big[V_i 1_{[|V_i|\leq T_n]}1_{[U_i\in\tilde{S}_1(u,u_k,h_n)]}\big] + \frac{C_{1,\mathsf{K}}^{-1}}{nh_n^m}\sum_{i=m}^{n-1}E\big[V_i 1_{[|V_i|\leq T_n]}1_{[U_i\in\tilde{S}_2(u,u_k,h_n)]}\big].\end{aligned}$$

To remove the dependence on u in these sums, we introduce

$$\Gamma_{k,n}^{(1)} := \bigcup_{u\in S_k\cap\mathsf{K}}\tilde{S}_1(u,u_k,h_n), \quad \Gamma_{k,n}^{(2)} := \bigcup_{u\in S_k\cap\mathsf{K}}\tilde{S}_2(u,u_k,h_n).$$

Then $\tilde{S}_1(u,u_k,h_n)\subset\Gamma_{k,n}^{(1)}$ and $\tilde{S}_2(u,u_k,h_n)\subset\Gamma_{k,n}^{(2)}$ for all $u\in S_k\cap\mathsf{K}$, and moreover

$$\lambda\left(\Gamma_{k,n}^{(1)}\right) = \lambda\left(\Gamma_{k,n}^{(2)}\right) \leq C_m h_n^{\gamma+m-1},$$

where C_m depends only on m. [For each $u\in S_k\cap\mathsf{K}$ the set $\tilde{S}_1(u,u_k,h_n)$ is one of the two disjoint components in the difference of two intersecting squares, and its area increases with the distance between u and u_k. The set $\Gamma_{k,n}^{(1)}$ is obtained by displacing u the farthest away from u_k in all directions; it is like a 'frame' to the square $S_1(u_k,h_n)$, and its measure is bounded above by a constant times $h_n^\gamma/2$ (the 'frame's width') times h_n^{m-1} (the 'frame's length').]

Therefore,

$$\sup_{u\in\mathsf{K}}\left|\tilde{\delta}_{1,n}(u)\right| \leq \max_{1\leq k\leq l_n}\frac{C_{1,\mathsf{K}}^{-1}}{nh_n^m}\sum_{j=1}^{2}\left\{\sum_{i=m}^{n-1}|V_i|1_{[|V_i|\leq T_n]}1_{[U_i\in\Gamma_{k,n}^{(j)}]} + \sum_{i=m}^{n-1}E\big[|V_i|1_{[|V_i|\leq T_n]}1_{[U_i\in\Gamma_{k,n}^{(j)}]}\big]\right\},$$

and we may finish the proof of (2.19) by showing that

$$\max_{1\leq k\leq l_n}\left|\frac{1}{nh_n^m}\sum_{i=m}^{n-1}|V_i|1_{[|V_i|\leq T_n]}1_{[U_i\in\Gamma_{k,n}^{(j)}]} - E\big[|V_i|1_{[|V_i|\leq T_n]}1_{[U_i\in\Gamma_{k,n}^{(j)}]}\big]\right| \to 0 \quad \text{a.s.} \tag{2.21}$$

and

$$\max_{1\leq k\leq l_n}\frac{1}{nh_n^m}\sum_{i=m}^{n-1}E\big[|V_i|1_{[|V_i|\leq T_n]}1_{[U_i\in\Gamma_{k,n}^{(j)}]}\big] \to 0. \tag{2.22}$$

The first statement follows just as $\max_{1\leq k\leq l_n}\delta_{1,n}(u_k)\to 0$ a.s. from an upper bound similar to (2.20), the only difference being that in the exponential term the factor h_n^m is replaced by $h_n^{m+1-\gamma}$. The second follows from the inequality

$$E\big[|V_i|1_{[|V_i|\leq T_n]}1_{[U_i\in\Gamma_{k,n}^{(j)}]}\big] \leq \int_{\Gamma_{k,n}^{(j)}}E[|V_i||U_i=u]dP(U_i\leq u) \leq ch_n^{\gamma+m-1}\max_{m\leq i\leq n-1}\sup_{u\in\mathsf{K}}E[|V_i||U_i=u],$$

where c is a constant, and assumption **A3**.

Note that in the course of proving **(c)** we have also shown (simply omit all the occurrences of the factor $|V_i|1_{[|V_i|\leq T_n]}$)

(d) $\sup_{u\in\mathsf{K}}\left|\Delta^*_{3,n}(u)\right|\to 0$ *a.s.*

To finish off we need to complete the proof of

(e) $\sup_{u\in\mathsf{K}}\left|\Delta^*_{1,n}(u)\right|\to 0$ *a.s.*
As in the proof of Theorem 2.2,

$$\begin{aligned}\sup_{u\in\mathsf{K}}\delta_{2,n}(u) &\leq \sup_{u\in\mathsf{K}}\frac{C_{2,\mathsf{K}}}{C_{1,\mathsf{K}}}\max_{m\leq i\leq n-1}\int_T^\infty P\left(|V_i|>v|U_i\in S(u,h_n)\right)dv\\ &\leq \left(\frac{C_{2,\mathsf{K}}}{C_{1,\mathsf{K}}}\right)^2\sup_{u\in\mathsf{K}}\int_{S(u,h_n)}\int_T^\infty 1-F(v|u')+F(-v|u)dv\frac{du'}{h_n^m},\end{aligned}$$

and by **(a)** this can be made arbitrarily small by taking T large enough.

Finally, we show that $\sup_{u\in\mathsf{K}}\delta_{3,n}(u)\to 0$ a.s. With $a_{n,i}(u)\equiv a_{n,i}$ and x_i as at the end of the proof of Theorem 2.2, we have

$$\delta_{3,n}(u)\leq\frac{C_{1,\mathsf{K}}^{-1}}{nh^n}\sum_{i=m}^{n-1}x_ia_{n,i}(u)\sum_{i=m}^{n-1}1_{[U_i\in S(u,h_n)]}\leq\left(\frac{C_{2,\mathsf{K}}}{C_{1,\mathsf{K}}}\sum_{i=m}^{n-1}x_ia_{n,i}(u)\right)\left|1+\Delta^*_{3,n}(u)\right|$$

By a slight generalization of Lemma 6.1.1 of Ash and Doléans-Dade (2000), it can be seen that this will tend uniformly to zero on K if $\sup_{u\in\mathsf{K}}a_{n,i}(u)\to 0$. But

$$\sup_{u\in\mathsf{K}}a_{n,i}(u)\leq\sup_{u\in\mathsf{K}}\frac{1_{[U_i\in S(u,h_n)]}}{\sum_{j=m}^{n-1}P(U_j\in S(u,h_n))}\sup_{u\in\mathsf{K}}\left|1+\Delta^*_{3,n}(u)\right|^{-1}\leq\frac{\sup_{u\in\mathsf{K}}\left|1+\Delta^*_{3,n}(u)\right|^{-1}}{C_{1,\mathsf{K}}nh_n^m}\to 0$$

because

$$\sup_{u\in\mathsf{K}}\frac{1}{\left|1+\Delta^*_{3,n}(u)\right|}=\frac{1}{\inf_{u\in\mathsf{K}}\left|1+\Delta^*_{3,n}(u)\right|}\leq\frac{1}{\inf_{u\in\mathsf{K}}1-\left|\Delta^*_{3,n}(u)\right|}=\frac{1}{1-\sup_{u\in\mathsf{K}}\left|\Delta^*_{3,n}(u)\right|}\to 1,$$

by **(d)**, so our proof is complete. □

3 Approximately Markov Sequences

There seems to be one main conceptual condition to be required from a process in order to be able to predict future from past values in a nonparametric way—by making use of its past observations only and without assuming much about its distribution. This condition, to which we may refer loosely as 'conditional stationarity', is that the same information at different moments in time yield the same probabilistic conclusions about the future. Without it, it seems that any prediction method would have to be capable of inferring or 'generalizing', which is

apparently impossible without postulating a model for the process. Mathematically, the idea may be formulated as follows: The process X is said to be *conditionally stationary* if there exist a function $(u, v) \to F(v|u)$ and m positive integers $i_1 < \cdots < i_m$ such that for all $x, x_1, \ldots, x_m \in \mathbb{R}$

$$P\left(X_{n+1} \leq x | X_{n-i_1+1} = x_m, \ldots, X_{n-i_m+1} = x_1\right) = F(x|x_m, \ldots, x_1). \tag{3.1}$$

(We have taken $i_j = j \ \forall j$ in our definition of Section 1.) Strictly stationary processes are conditionally stationary, and so are homogeneous Markov processes of arbitrary order; on the other hand, inhomogeneous Markov processes are not conditionally stationary. Processes with stationary, non-independent increments, such as fractional Brownian motion with $\mathbb{N}$ as index set, provide other examples of processes which are not conditionally stationary.

The definition of a conditionally stationary process goes hand in hand with this forecasting principle: fix a sub-path conditional upon which predictions are to be made, look back at the realization of the process in search of sub-paths which somehow resemble it, and then use the knowledge of how these sub-paths have evolved to guess the next value of the process. In the form of estimation of conditional distribution functions and expectations, this principle is plainly justified because it leads to consistent estimators, at least for conditionally stationary processes satisfying assumptions similar to those of Section 2. From the point of view of applications it can be justified on empirical grounds for processes which to some extent may be regarded as conditionally stationary; the statistical analysis of Section 5 illustrates how graphical and statistical tools might be used to check the conditional stationarity hypothesis.

Of course, prediction methods based on such a principle are generally sub-optimal because they involve conditioning on a portion of the past rather than on the full observed path. In itself this may not be a problem at all, especially when a parametric method is not tenable. Sometimes, however, we would like to say a bit more about what may be expected from nonparametric estimators such as $R_n(X_n, \ldots, X_{n-m+1})$ (i.e., (1.4) evaluated at the point $u = (X_n, \ldots, X_{n-m+1})$), in particular when we want to compare them with parametric estimators in cases where the latter are optimal. One may well wonder whether the errors of a nonparametric predictor can be brought near those of the optimal predictor, whether they do not increase with time, etc. To proceed in this direction we shall need to strengthen (3.1) by assuming that there exists an $m_0 \in \mathbb{N}$ such that for all integers $n \geq m \geq m_0$ and all $x, x_m, \ldots, x_1 \in \mathbb{R}$

$$P\left(X_{n+1} \leq x | (X_n, \ldots, X_{n-m+1}) = (x_m, \ldots, x_1)\right) = F(x|x_m, \ldots, x_1). \tag{3.2}$$

Under this assumption the prediction principle can be carried out by conditioning on paths $(X_{i-m+1}, \ldots, X_i)$, $i = m, \ldots, n-1$, of any length $m \in [m_0, n]$, and one may more specifically ask whether the prediction of a future observation on the basis of a fixed number of past observations is 'good enough', or if instead the number of observations used to produce forecasts should somehow increase with the sample size. On a practical level the latter option appears to be extremely undesirable because any specification of the rate at which the number of observations used to forecast grows would certainly demand a considerable knowledge of the process, and even if such rate were available it would lead to limit theorems in terms of vectors of increasing dimension, whose practical interpretation is often hard to grasp.

Prediction based on a fixed number of past observations is much more practical and easily interpretable, and it would be interesting to provide it with some sort of theoretical justification.

With this in mind, let us try and focus on processes for which, roughly speaking, the error incurred when predicting in the sub-optimal way can be kept small and be non-increasing with the sample size. If the interest lies in estimators of the conditional expectation of X_{n+1} given $X_n, \ldots, X_1$, this can be translated into the requirement that, given a positive integer $m \in [m_0, n]$, the error

$$\delta_{n,m}(X_n, \ldots, X_1) := \left| \Pi_n^{(n)} - \Pi_m^{(n)} \right|,$$

where $\Pi_m^{(n)} \equiv \Pi_m^{(n)}(X_n, \ldots, X_{n-m+1}) := E[X_{n+1}|X_n, \ldots, X_{n-m+1}]$ for $1 \leq m \leq n$, be, in some sense, kept smaller than a certain $\delta > 0$ *for all* $n \geq m$.

If X is a Markov process of order m, then $\delta_{n,m}(X_n, \ldots, X_1) = 0$; although the event $[\delta_{n,m}(X_n, \ldots, X_1) < \delta]$ does not generally hold with probability 1, it typically holds with high probability if m is large enough, which suggests this definition for the class of processes we are thinking of: A process X is said to be *approximately Markov* (*of order* m_0) if m_0 is the smallest integer for which (3.2) is satisfied and if for any $\varepsilon, \delta > 0$

$$\sup_{n \geq m} P\left(\delta_{n,m}(X_n, \ldots, X_1) > \delta \right) \leq \varepsilon \tag{3.3}$$

for sufficiently large $m \geq m_0$.

To illustrate the sort of statements about $R_n(X_n, \ldots, X_{n-m+1})$ this definition permits, suppose X is approximately Markov, that for a fixed m for which (3.3) holds we have $\sup_{u \in \mathsf{K}} |R_n(u) - R(u)| \to^P 0$ for every compact set $\mathsf{K} \subset \mathbb{R}^m$, as we do for example under conditions like those of Theorem 2.3, and furthermore that the sequence $\{U_i\}$ is tight. What we can say, then, is that *given arbitrary* $\varepsilon, \delta > 0$ *we can fix* $m \in \mathbb{N}$ *such that*

$$P\left(\left| R_n(X_n, \ldots, X_{n-m+1}) - \Pi_n^{(n)}(X_n, \ldots, X_1) \right| > \delta \right) \leq \varepsilon \tag{3.4}$$

for all sufficiently large n. This is precisely the conclusion one would like to draw, at least in one of its forms: for large enough n, the number of past observations m used to estimate the expectation of the future observation conditional on *the whole* observed path does not have to increase with the sample size in order to keep the error of the nonparametric predictor $R_n(X_n, \ldots, X_{n-m+1})$ relative to the optimal predictor $\Pi_n^{(n)}$ smaller than a prescribed $\varepsilon > 0$.[4]

This result can be used in many situations to justify prediction based on a fixed number of past observations; the simulation study of Section 4 will illustrate just this in the case of a Gaussian ARMA$(1,1)$ process.

We close this section with an example of a class of (stationary) approximately Markov processes. Let X be a stationary Gaussian process with covariance function γ satisfying $\gamma(0) > 0$ and $\gamma(k) \to 0$ as $k \to \infty$. For each positive integer $m \leq n$, the covariance matrix of the vector $(X_n, \ldots, X_{n-m+1})$, which we designate by Σ_m, is non-singular (e.g. Brockwell and Davis (1987), p. 160). Write, for $1 \leq m \leq n$,

$$\pi_m\left(x_n, \ldots, x_{n-m+1}\right) = \gamma_m \Sigma_m^{-1} \mathbf{x}_{n,m}^T, \quad (x_n, \ldots, x_{n-m+1}) \in \mathbb{R}^m,$$

[4]Of course, the existence of moments of order $r > 1$ implies a similar statement in terms of expected values.

where $\gamma_m = [\gamma(1) \cdots \gamma(m)]$, $\mathbf{x}_{n,m} = [x_n \cdots x_{n-m+1}]$. As is well known,

$$\Pi_m^{(n)} = E\left[X_{n+1} \,|X_n, \ldots, X_{n-m+1}\right] = \pi_m\left(X_n, \ldots, X_{n-m+1}\right)$$

and

$$E\big[(\Pi_m^{(n)} - X_{n+1})^2\big] = \gamma(0) - \gamma_m \Sigma_m^{-1} \gamma_m^T \geq \gamma(0) - \gamma_n \Sigma_n^{-1} \gamma_n^T = E\big[(\Pi_n^{(n)} - X_{n+1})^2\big].$$

Since $\sigma_n^2 := \gamma(0) - \gamma_n \Sigma_n^{-1} \gamma_n^T$ decreases with n, $\sigma^2 := \lim_n \sigma_n^2$ exists. Also, $E\big[(\Pi_m^{(n)} - \Pi_n^{(n)})^2\big] = \sigma_m^2 + \sigma_n^2 + 2E\big[(\Pi_n^{(n)} - X_{n+1})X_{n+1}\big] + 0 = \sigma_m^2 + \sigma_n^2 + 2\gamma_n \Sigma_n^{-1} \gamma_n^T - 2\gamma(0) = \sigma_m^2 - \sigma_n^2$. Thus, given arbitrary $\varepsilon > 0$,

$$E\big[(\Pi_m^{(n)} - \Pi_n^{(n)})^2\big] \leq \sigma_m^2 - \sigma^2 < \varepsilon$$

for all $n \geq m$ if m is large enough, and this implies (3.3).

A fact worth mentioning in this connection is contained in Section 10.10 of Grenander and Szegö (1958): if X has an analytic spectral density with no real zeros and integrable logarithm—hence, in particular, if X is a causal and invertible ARMA process—then $\sigma_m^2 - \sigma^2 \to 0$ as $m \to \infty$ exponentially fast. Because (3.4) can only benefit from such speed of convergence, this explains why in many applications the number m of past observations required for a satisfactory prediction is surprisingly small. For the ARMA$(1,1)$ process considered next one can compute $\sigma_m^2 - \sigma^2$ explicitly.

4 An Illustration with a Gaussian ARMA(1,1) Process

We consider the process defined by

$$X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1}, \quad t \in \mathbb{N},$$

where $\{Z_t : t \in \mathbb{Z}\}$ is an i.i.d. $N(0, \sigma^2)$ sequence and $|\phi|, |\theta| < 1$. If we assume that $(1 - \phi z)(1 + \theta z) \neq 0$ for $|z| \leq 1$ then the process is causal and invertible and, by Theorem 9 on p. 189 of Ibragimov and Rozanov (1978) and Theorem 4.4.2 of Brockwell and Davis, the mixing property (2.1) holds with $\limsup_{k \to \infty} \alpha(k)^{1/k} \leq |\phi|$.

For this process one can compute explicitly several characteristics of interest (Example 5.3.3 of Brockwell and Davis (1987) contains some of the formulae below). The covariance function is given by

$$\gamma(k) \equiv E[X_t X_{t+k}] = \begin{cases} \sigma^2 \left(1 + 2\,\phi\,\theta + \theta^2\right) / \left(1 - \phi^2\right) & \text{if } k = 0, \\ \sigma^2 \phi^{k-1} \left(\phi + \theta + \phi^2\theta + \phi\,\theta^2\right) / \left(1 - \phi^2\right) & \text{if } k = 1, 2, \ldots \end{cases}$$

The minimum mean square error (MSE) predictor of X_{n+1} based on $X_n, \ldots, X_1$ can be recursively calculated as $\Pi_0 \equiv 0$,

$$\Pi_n^{(n)} = \phi X_n + \frac{\theta}{r_{n-1}}(X_n - \Pi_{n-1}), \quad n = 1, 2, \ldots, \tag{4.1}$$

where

$$r_n = 1 + \theta^{2n} \left(\frac{1 - \phi^2}{(\theta + \phi)^2} + \frac{1 - \theta^{2n}}{1 - \theta^2} \right)^{-1}$$

for $n = 1, 2, \ldots$, and $r_0 = \gamma(0)/\sigma^2$. Setting $p_0 = 0$ and

$$p_i = \phi X_{i+n-m} + \frac{\theta}{r_{i-1}}(X_{i+n-m} - p_{i-1}), \quad i = 1, \ldots, m,$$

we can use (4.1) to conclude that $\Pi_m^{(n)} = p_m$ is the minimum MSE predictor of X_{n+1} based on $X_n, \ldots, X_{n-m+1}$. Finally, the conditional distribution of X_{n+1} given $X_n, \ldots, X_1$ is normal with mean $\Pi_n^{(n)}$ and variance

$$E\big[(X_{n+1} - \Pi_n^{(n)})^2\big] = \sigma_n^2 = \sigma^2 r_n.$$

By what was said in Section 3 the process is approximately Markov, and we have

$$E\big[(\Pi_m^{(n)} - \Pi_n^{(n)})^2\big] = E\big[(\Pi_m^{(n)} - X_{n+1})^2\big] - E\big[(\Pi_n^{(n)} - X_{n+1})^2\big] \le \sigma^2(r_m - 1) \tag{4.2}$$

for all $n \ge m$, which gives an upper bound for the mean square difference (MSD) between $\Pi_m^{(n)}$ and $\Pi_n^{(n)}$.

Table 1: *Upper bounds for the mean square difference (MSD) between* $\Pi_n^{(n)}$ *and* $\Pi_m^{(n)}$ *as given in (4.2) corresponding to a choice of the parameters* ϕ *and* θ *of an ARMA*$(1,1)$ *model with* $\sigma^2 = 1$ *and of the number* m *of past observations.*

ϕ	θ	m	Bound MSD
0.1	0.4	1	0.0320
0.1	0.4	2	0.0050
0.7	0.7	2	0.1372
0.7	0.7	3	0.0591
0.7	0.7	4	0.0273
0.7	0.7	5	0.0130
0.7	0.7	6	0.0063

Table 1 shows values of the right-hand side of (4.2) corresponding to choices of m and of the parameters of the ARMA$(1,1)$ model with $\sigma^2 = 1$. The figures give an idea of what kind of errors can be expected when forecasting using the previous m observations in place of all past observations. When $\phi = 0.1$ and $\theta = 0.4$ we have $\gamma(0)^{1/2} \approx 1.21$, so the improvement of $\Pi_n^{(n)}$ relative to $\Pi_1^{(n)}$ is at most about 15% of the standard deviation of the time series. When $\phi = \theta = 0.7$ we have $\gamma(0)^{1/2} \approx 2.20$, whence the improvement of $\Pi_n^{(n)}$ relative to $\Pi_2^{(n)}$ is at most 17% of the standard deviation of the time series. Thus, even in 'high correlation' cases a relatively small m will already produce practically satisfactory point estimates (a mere reflection of the exponential decay of the MSD). By taking $m = 2$ and $m = 5$ in the first and second cases considered, one can further reduce the improvement of $\Pi_n^{(n)}$ upon $\Pi_m^{(n)}$ to about 5% of the corresponding standard deviations.

Table 1 gives upper bounds for *parametric* errors. The sub-optimality of our prediction method results from parametric and *nonparametric* errors—those arising from the nonparametric

estimation of quantities based on the distribution of $\Pi_m^{(n)}$. For example, in terms of MSE we have

$$E[(\tilde{\Pi}_m^{(n)} - X_{n+1})^2] = E[(\Pi_m^{(n)} - X_{n+1})^2] + E[(\tilde{\Pi}_m^{(n)} - \Pi_m^{(n)})^2],$$

where

$$\tilde{\Pi}_m^{(n)} := R_n(X_n, \ldots, X_{n-m+1});$$

in other words, the error incurred using the nonparametric method is the sum of parametric and nonparametric errors. Since there should be a trade-off between the two types of errors (smaller parametric errors implying larger nonparametric errors, and conversely), it is of some interest to see what form this trade-off may assume in practice, and what its consequences are as regards pointwise and interval prediction.

With this in mind we have used simulation to compute approximately some characteristics of the ARMA$(1,1)$ model with $\phi = \theta = 0.7$ and $\sigma^2 = 1$. Table 2 shows estimates of the MSE of the predictors $\Pi_m^{(n)}$ and $\tilde{\Pi}_m^{(n)}$—MSE$(\Pi_m^{(n)})$, etc., for short—and coverage probabilities of the approximate 95% predictive intervals defined by (2.11), for $m = 2, 6$ and sample sizes of $n = 125 \times 2^0, \ldots, 125 \times 2^{13}$. The results are based on 100,000 simulations, whence the widths of the 95% confidence intervals for the MSE and coverage probability estimates are less than 0.02 and 0.004, respectively.

Of course, MSE$(\Pi_m^{(n)})$ is known exactly; we have included its estimates both as a check and as a means of illustrating the adherence of (4.2).

Table 1 gives us 0.1372 as an upper bound for the MSD between $\Pi_n^{(n)}$ and $\Pi_m^{(n)}$; the results of Table 2 for $m = 2$ are consistent with this for all sample sizes. Of course, within the accuracy of the table MSE$(\Pi_m^{(n)})$ is not supposed to decrease with n, but MSE$(\tilde{\Pi}_m^{(n)})$ is. For $m = 2$, MSE$(\tilde{\Pi}_m^{(n)})$ is at the beginning about 50% larger than the asymptotic error of $\Pi_n^{(n)}$, and 38% larger than MSE$(\Pi_m^{(n)})$; it decreases and practically attains MSE$(\Pi_m^{(n)})$ when $n = 1,024.000$.

For $m = 6$ the bound for the MSD given in Table 1 is < 0.007; accordingly, the difference between the asymptotic error of $\Pi_n^{(n)}$ and the MSE of $\Pi_m^{(n)}$ in Table 2 is practically insignificant. The convergence of MSE$(\tilde{\Pi}_m^{(n)})$ to MSE$(\Pi_m^{(n)})$ is in this case extremely slow: for small n, MSE$(\tilde{\Pi}_m^{(n)})$ can exceed MSE$(\Pi_m^{(n)})$ (and MSE$(\Pi_n^{(n)})$) by as much as 140%, and for $n = 128,000$ by 25%; in contrast, an excess of 25% over MSE$(\Pi_m^{(n)})$ occurs in the case $m = 2$ only for $n < 1000$. Less accurate simulations of larger samples indicate that MSE$(\tilde{\Pi}_m^{(n)})$ attains MSE$(\Pi_m^{(n)})$ to within the accuracy of Table 1 when $n = 8,192,000 = 8 \times 1,024,000$. Thus, if we take into account the fact that the standard deviation of this ARMA$(1,1)$ model is $\gamma(0)^{1/2} \approx 2.20$, we see that from a MSE point of view the price to pay for a gain in precision of $0.13 \approx 0.1372 - 0.0063$, achieved by predicting with $\tilde{\Pi}_6$ in place of $\tilde{\Pi}_2$, is extremely high.

With both $m = 2$ and $m = 6$ the predictive intervals are too narrow for small n, and with $m = 6$ they are slightly conservative for large n, the convergence to the true coverage probability being more rapid when $m = 2$. In any case, the results show that one can make practically reliable predictive statements based on the empirical conditional distribution function with rather small sample sizes—$n = 2000$, say—even for relatively large m, which is in contrast with the results about conditional means.

We should say something about the choice of the 'smoothing parameter' h_n. For exponential decreasing mixing coefficients we require $n^{1-1/r}h_n^m/(\log n)^2 \to \infty$ for a given $r > 1$, which suggests taking $h_n = c_m \left(n^{\gamma-1}(\log n)^2\right)^{1/m}$ with $\gamma \in (0,1)$ and c_m to be determined. By minimizing the prediction error in terms of the supremum distance between the empirical and true conditional distribution functions, we have carried out some simulation experiments to numerically determine optimal values of h_n for various sample sizes, and then determined γ and c_m by fitting the mapping $n \to c_m \left(n^{\gamma-1}(\log n)^2\right)^{1/m}$ to those values of h_n. For $m = 2$ this procedure gave $\gamma = 0.45$, $c_m \approx 1.32$, and for $m = 6$ it gave $\gamma = 0.01$, $c_m \approx 5.83$. These settings are far from being critical; other choices of distances and functional forms for $n \to h_n$ yield neighbouring values for the parameters and very similar results.

Table 2: *Estimates based on 100,000 simulations of the ARMA*$(1,1)$ *model with* $\sigma^2 = 1$ *and* $\phi = \theta = 0.7$ *of the mean square errors of the predictors* $\Pi_m^{(n)}$ *(parametric) and* $\tilde{\Pi}_m^{(n)}$ *(nonparametric) and of the coverage probabilities of the* 95% *nominal predictive intervals based on the empirical conditional distribution function (see (2.11)) for various sample sizes,* $m = 2$ *and* $m = 6$.

	$m = 2$			$m = 6$		
n	MSE($\Pi_m^{(n)}$)	MSE($\tilde{\Pi}_m^{(n)}$)	Cov. prob.	MSE($\Pi_m^{(n)}$)	MSE($\tilde{\Pi}_m^{(n)}$)	Cov. prob.
125	1.136	1.562	0.793	1.002	2.394	0.831
250	1.137	1.390	0.864	0.997	2.044	0.889
500	1.145	1.298	0.902	1.013	1.868	0.918
1000	1.133	1.232	0.924	1.010	1.743	0.937
2000	1.146	1.212	0.936	1.001	1.611	0.948
4000	1.136	1.179	0.943	0.993	1.515	0.955
8000	1.138	1.165	0.948	1.006	1.464	0.959
16,000	1.138	1.155	0.950	1.007	1.399	0.962
32,000	1.141	1.152	0.952	1.004	1.350	0.964
64,000	1.137	1.144	0.952	1.012	1.304	0.964
128,000	1.137	1.143	0.953	0.998	1.254	0.965
256,000	1.123	1.126	0.953	1.006	1.219	0.964
512,000	1.139	1.141	0.951	1.006	1.184	0.964
1,024,000	1.135	1.136	0.952	1.002	1.152	0.964

To sum up our conclusions, let us say that for this type of model our nonparametric pointwise estimates and prediction intervals with small m are good enough for most practical purposes one can think of, even for relatively small samples, and that a 'too accurate' estimation demands rather large sample sizes—of the order of millions, really. What is also important to observe is how the numerical results do indeed illustrate the argument around (3.4): once specific precision requirements are made—e.g. MSD≤ 0.1372—one can actually see them being fulfilled after some time. Thus one may well say that for large samples the nonparametric method works in a nearly optimal way—almost in the same way as if the true model were completely known.

This numerical study seems relevant to those scientific areas dealing with environmental data, where the time series can often be assumed approximately Markov of rather low order (from which stems the 'difficulty' in predicting them); the application described in the next section provides an example of this situation.

5 An Illustration with Wave Data

In Ocean Engineering, the sea surface elevation relative to the mean water level at a given point and during a period of about 3 hours is modelled as a stationary, ergodic, mean zero Gaussian process. This model is known to provide a reasonable approximation to wave phenomena, and it enables the calculation of several *sea-state parameters* which serve to partly characterize the sea-state. The evolution of the sea-state in time is regarded as a sequence of such Gaussian processes, and accordingly can be described in terms of time series of sea-state parameters.

One of the most useful of these parameters is the so-called *significant wave height* (SWH), which is defined as 4 times the standard deviation of the sea surface elevation process and hence provides a measure of severity of the sea-state. Naturally, the prediction of future values of SWH is a matter of importance in Ocean Engineering and in the oceanographic sciences in general. Currently, forecasts are provided by state-of-the-art *wave models*, models based on the physical description of wave generation, dissipation and non-linear wave-wave interactions, that use wind fields, past wave observations, and even *present* wave observations from certain geographical positions (a technique called 'assimilation'), to produce estimates of wave conditions on a grid of points over whole ocean areas.

Here we shall use series of SWH buoy measurements from the American National Oceanic and Atmospheric Administration (NOAA) database[5] to study the quality of 1-step forecasts of SWH based on nonparametric estimators of the regression and conditional distribution functions. The results will be compared in terms of mean square error with the corresponding wave model predictions from the ERA-40 project, a project aiming at the reconstruction of global wave conditions from 1957 onwards (see Caires and Sterl (2003)).

The dataset we have chosen consists of the time series of 3-hourly averages of SWH (in metres) at a location off the northeast coast of the U.S.A. (measurements from buoy 46006) from 1978 to 1996. The 3-hourly averages provide a description of the severity of the sequence of sea-states postulated in the Ocean Engineering model. We shall construct estimators of the regression and conditional distribution functions using the data from 1978 to 1995, comprising a total of 18535 measurements (there are a number of gaps in the series), and use them to produce one-step predictions for May 1996. Typically, the month of May includes a great variety of wave conditions, including *wind sea*—locally generated waves—and *swell*—travelling waves generated elsewhere—, and that is why we have chosen it for our illustrations.

The left panel of Figure 1 shows the time series of 3-hourly averages of SWH during May 1996. The non-stationary character of the process seems evident from the figure. On the other hand, both experimental and theoretical physics point to general laws by which waves evolve from a given set of initial conditions, which suggests that the assumption of a conditional law

[5]Available at http://seaboard.ndbc.noaa.gov/

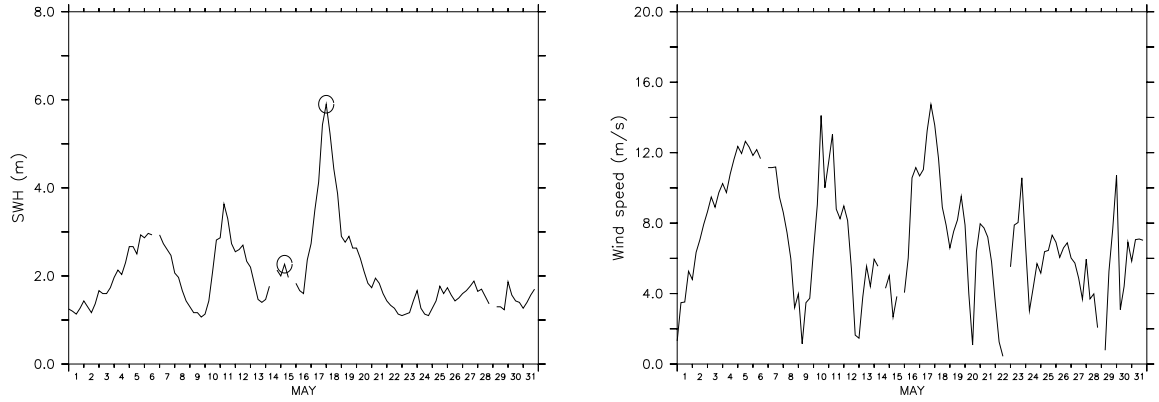

Figure 1: Wave and wind observations during May 1996 from a buoy located in the Northeastern Pacific. Left panel: significant wave height; right panel: wind speed.

for one observation given a few previous observations might not be an unreasonable one.

The right panel of the figure shows the time series of wind speed (at a 10-metre height, in metres per second) at the buoy location. The overall correspondence between the wind and wave time series reflects the fact that waves are generated by the wind. However, the relationship between winds and waves is very complex; it depends on many factors and to a certain extent has to be treated as random. For instance, the first two peaks in the time series of SWH shown in Figure 1 can be explained in terms of the corresponding peaks in the time series of wind speed, but the occurrence of the highest peak in the time series of SWH must be attributed to a combination of wind sea and swell.

Figure 2 shows one-step forecasts for May 1996 obtained from the empirical regression function, together with the time series of buoy measurements, the envelope representing the endpoints of the 95% predictive intervals computed from the empirical conditional distribution function, and the sequence of errors (the differences between measurements and forecasts). The predictions are based on $m = 2$ past observations; numerical studies with the data from 1978 to 1995 indicate that this is likely to be the best choice in terms of mean-square error, besides being in agreement with the 'memory length' expected from physical models and empirical evidence. The choice of the smoothing parameter was carried out essentially as in the case of the simulations of Section 4 (using only the data from 1978 to 1995 and not that from 1996) but aiming at the minimization of the mean-square error.

From the 107 forecasts made (a few measurements were missing) we compute an average error or bias of 0.01 metres, a root mean square error (square root of the sum of squared errors) of 0.31 metres, and a coverage rate of the predictive intervals of 98%. As expected, the latter are above the prescribed 95% because of the dependence between observations. The corresponding forecasts furnished by the above mentioned ERA-40 wave model have a bias of -0.15 metres and a root-mean square error of 0.34 metres; wave model results give no confidence intervals.

In the way they are used here the non-parametric estimators do not offer an alternative

to global wave models, since the latter provide forecasts over whole areas rather than at single location, but it is clear from these results that they can be profitably used at least for local studies.

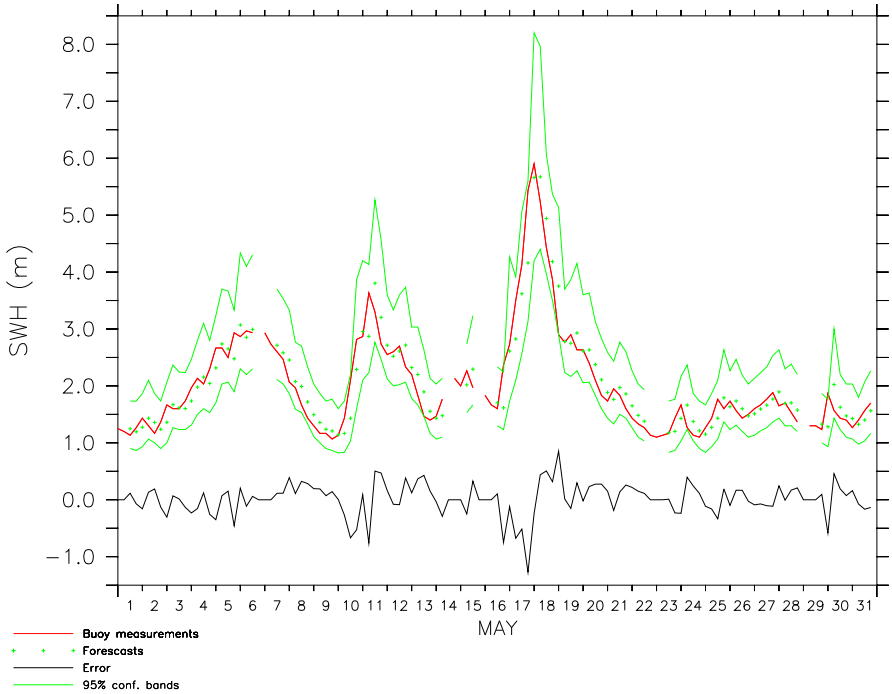

Figure 2: One-step predictions and SWH measurements for May 1996, and the corresponding 95% confidence bands and errors.

Besides being useful in predicting wave conditions, nonparametric estimators are also potentially useful in modelling wave dynamics and in studying climate changes. Indeed, assuming there are approximate physical laws governing the evolution of the sea state given a set of initial conditions, the empirical conditional distribution function of a (future) sea-state parameter given the values of other parameters (eg. past or neighbouring observations from one or more variables) represents the ideal tool to infer and validate such laws; on the other hand, the occurrence of climate changes might be detected through systematic changes in conditional laws or conditional means. The remainder of this section will serve to illustrate these points.

Table 3 shows the biases, root-mean-square errors and coverage probabilities of the predictive intervals with nominal level 0.95 of the SWH forecasts for May 96 based on various subsets of the 1978-1995 data set, namely the subsets of 1, 2, 3, 6, 9 years and the full 18-year set. The root mean square error decreases by about 8% as the amount of data increases from 1 to 9 years, but there is no significant improvement when passing from 9 to 18 years of data, indicating that with only two past observations smaller errors should not be expected (even from an extremely good parametric model) and that a past history of 9000 observations is practically enough to reach the ideal situation.

Table 3: Biases, root-mean-square errors (RMSE) and coverage probabilities of the predictive intervals of the SWH forecasts for May 96 based on different subsets of the 1978-1995 database.

Amount of data	Period	n	Bias	RMSE	Cov. prob.
1 year	1978–1978	1116	-0.020	0.334	0.981
	1979–1979	1132	-0.016	0.338	1.000
	1979–1979	1132	-0.016	0.338	1.000
	1982–1982	1447	-0.005	0.336	0.991
	1983–1983	1428	-0.009	0.341	0.991
	1984–1984	1458	0.005	0.339	1.000
	1985–1985	1200	0.054	0.336	0.991
	1987–1987	1199	0.046	0.340	0.991
	1989–1989	1053	0.054	0.347	0.981
	1990–1990	1430	-0.011	0.344	0.953
	1993–1993	1448	0.023	0.334	0.972
	1994–1994	1129	0.015	0.332	0.981
2 years	1978–1979	2249	-0.016	0.322	0.991
	1980–1981	940	-0.029	0.358	0.963
	1982–1983	2878	-0.003	0.328	1.000
	1984–1985	2661	0.017	0.316	1.000
	1986–1987	2166	0.014	0.324	0.963
	1988–1989	1928	0.041	0.326	0.991
	1990–1991	1461	-0.007	0.342	0.963
	1992–1993	2234	0.015	0.322	0.972
	1994–1995	1995	0.023	0.327	0.991
3 years	1978–1980	2249	-0.016	0.322	0.991
	1981–1983	3821	-0.008	0.337	0.972
	1984–1986	3628	0.012	0.311	0.991
	1987–1989	3130	0.040	0.324	0.981
	1990–1992	2245	0.001	0.331	0.972
	1993–1995	3446	0.023	0.317	0.972
6 years	1978–1983	6070	-0.011	0.323	0.972
	1984–1989	6761	0.024	0.311	0.981
	1990–1995	5694	0.017	0.316	0.963
9 years	1978–1986	9701	0.001	0.311	0.991
	1987–1995	8827	0.028	0.312	0.953
18 years	1978–1995	18531	0.014	0.309	0.981

The biases show that the nonparametric estimators are almost unbiased for small as well as large sample sizes. The predictive intervals are always conservative, though less so when bigger samples are used, showing that the excessive coverage is mainly a consequence of the dependence in the sequence of forecasts.

Observe that the results are rather consistent within sample size categories; in particular, the nonparametric estimators constructed with the first and second 9-year periods are very similar as far as the statistics shown in Table 3 are concerned. This supports the hypothesis of conditional stationarity and suggests no climate changes have occurred in this location *at least at the level of conditional means*.

Table 4: Mean- and median-based forecasts of typical and atypical observations using three different data sets, with the corresponding 95% predictive intervals ($\hat{x}_{0.025}$ and $\hat{x}_{0.975}$) and the numbers of observations (No. obs.) contributing to the nonparametric estimators.

Predicted observation	Period	No. obs.	Mean	Median	$\hat{x}_{0.025}$	$\hat{x}_{0.975}$
15/5/1996, 00h:00m	1978-1986	847	2.037	1.967	1.533	2.933
2.267 m	1987-1995	844	2.009	2.000	1.567	2.700
	1978-1995	1323	2.018	1.967	1.533	2.733
18/5/1996 06h:00m	1978-1986	17	5.457	5.433	4.467	6.833
5.9 m	1987-1995	15	5.898	5.533	4.200	8.200
	1978-1995	27	5.659	5.433	4.2	8.200

To take a closer look at the conditional stationarity assumption we shall now concentrate on two observations of May 1996, one that is rather typical, of 2.267 m., and another more atypical or extreme, of 5.9 m (the largest observation in the series); they are indicated by circles in the left panel of Figure 1. Table 4 shows forecasts of these observations based on the conditional mean and conditional median as well as the endpoints of the corresponding 95% predictive intervals and the number of observations involved in the construction of the empirical regression and conditional distribution functions—i.e., the number of strictly positive terms in (1.3) and (1.4). Three forecasts are given for each observation: one is obtained using the data from the first 9-year period (1978-1986), the other the data from the second 9-year period (1987-1995), and the third the whole 18-year series. As one would expect, the number of observations contributing to the nonparametric estimators is quite large in the case of the typical observation but rather small in the case of the extreme observation.

Let us note that the ERA-40 wave model underestimates the extreme observation by about 1.5 metres; in contrast, the nonparametric estimators are able to capture the quick increase and uncertainty in the SWH values; both means and medians have a rather small error, and the actual observation falls comfortably into each of the three predictive intervals.

The predictive intervals for the typical observation indicate a good agreement between the estimators obtained from the first and second halves of the data, but perhaps the same can not be said of the intervals for the extreme observation. Figures 3 and 4 allow an overall comparison between the empirical conditional distributions obtained from the two 9-year datasets and

between each of these and the empirical conditional distribution constructed from the whole 18-year series; they show quantile plots (the quantiles of one empirical distribution function versus the quantiles of another) and graphs of the empirical conditional distribution functions.

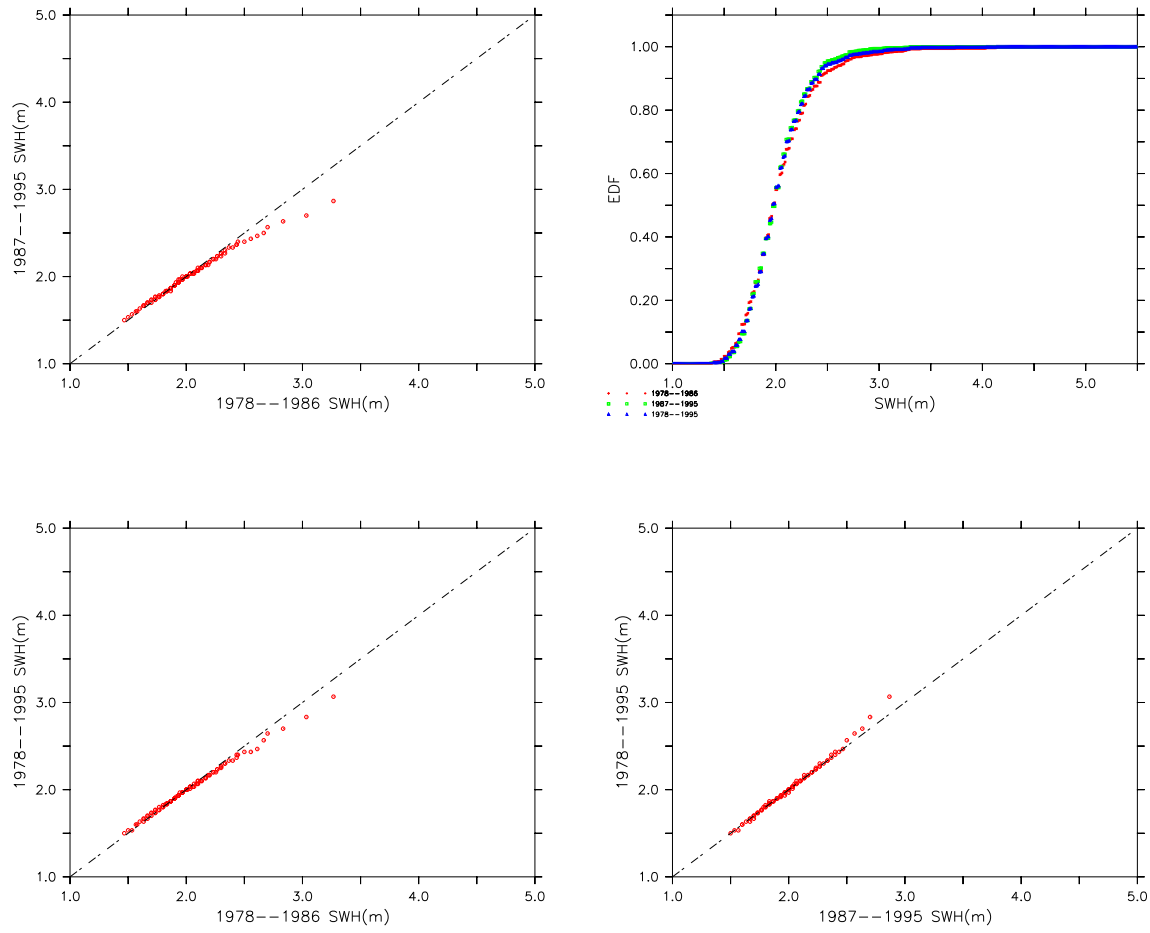

Figure 3: Graphs comparing the empirical conditional distributions of the typical observation constructed from three different data sets; upper left and lower panels: quantile plots; upper right panel: empirical distribution functions.

Figure 3 reveals a good agreement between the three estimators in the range of 0 to 2.5 metres but a discrepancy above 2.5 metres. Unfortunately, the dependent character of the data makes it difficult to test whether the empirical distributions from the two halves of the data are significantly different or not; a two-sample Kolmogorov-Smirnov test gives a p-value of about 0.55, but there is no reason why this should be taken into account and one can only conclude that in spite of some agreement between the two distributions the hypothesis of conditional stationarity may represent only a crude approximation.

The plots in Figure 4 indicate a reasonable agreement between the two empirical conditional distributions for the atypical observation, but there is again evidence of differences in the upper tails: the distribution function constructed from the second half of the data has a comparatively heavier tail (a consequence of the presence of certain more extreme events observed during 1987-1995). The two-sample Kolmogorov-Smirnov test gives a p-value of 0.80 (corresponding to a maximal distance of 0.038 between the two distribution functions); since in this case the observations contributing to the estimators are rather rare and spaced, and consequently approximately uncorrelated, the non significant result of the test deserves to be taken seriously.

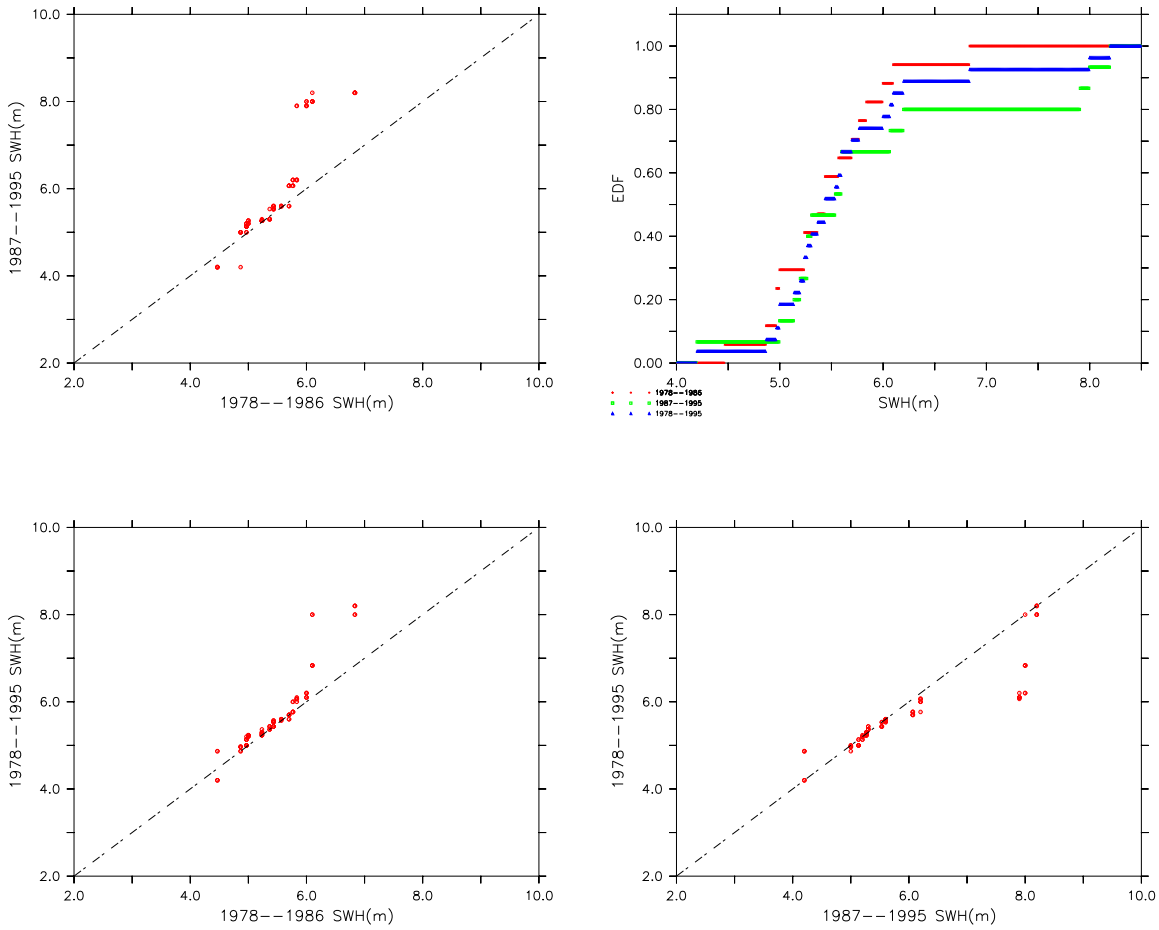

Figure 4: Graphs comparing the empirical conditional distributions of the atypical observation constructed from three different data sets; upper left and lower panels: quantile plots; upper right panel: empirical distribution functions.

In conclusion, one may state that in spite of some discrepancies between the empirical conditional distribution functions computed from data of different periods the conditional stationarity assumption seems to be fulfilled at least approximately in the case of this SWH dataset. Of course, our somewhat subjective considerations can not replace formal (and more global) testing procedures, but the latter seem difficult to obtain in so general a setting.

Aknowledgements
The second author is grateful to Enrico Capobianco and Kacha Dzhaparidze for useful conversations on this topic. We thank the American National Oceanographic Data Center for the buoy data.

References

Ango Nze, P., and Doukhan, P. (2002). Weak dependence: models and applications. In H. Dehling, T. Mikösh, and M. Sorensen (Eds.), *Empirical Process Techniques for Dependent Data* (p. 117-136). Birkhäuser.

Ash, R., and Doléans-Dade, C. (2000). *Probability and Measure Theory* (2 ed.). Academic Press.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes* (2 ed.). Lecture Notes in Statistics 110, Springer-Verlag.

Brockwell, P., and Davis, R. (1987). *Time Series: Theory and Methods.* Springer-Verlag.

Caires, S., and Sterl, A. (2003). Validation of ocean wind and wave data using triple collocation. *J. Geophys. Res.*, *108*(C3), 43.1-43.16, doi:10.1029/2002JC001491.

Carbon, M. (1983). Inégalité de Bernstein pour les processus fortement mélangeants, non nécessairement stationaires. Applications. *C. R. Acad. Sc. Paris*, *I, t. 297*, 303-306.

Collomb, G. (1984). Propriétés de convergence presque complète du prédicteur à noyau. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, *66*, 441-460.

Collomb, G. (1985). Nonparametric regression: an up-to-date bibliography. *Statistics*, *16*, 309-324.

Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *Annals of Statistics*, *9*(6), 1310-1319.

Grenander, U., and Szegö, G. (1958). *Toeplitz Forms and their Applications.* University of California Press.

Györfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series.* Lecture Notes in Statistics 60, Springer-verlag.

Härdle, W. (1989). *Applied Nonparametric Regression.* Cambridge: Cambridge University Press.

Ibragimov, I., and Rozanov, Y. (1978). *Gaussian Random Processes.* Springer-Verlag.

Knopp, K. (1928). *Theory and Application of Infinite Series.* London and Glasgow: Blackie & Son.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and its Application*, *9*, 141-142.

Roussas, G. (1969). Nonparametric estimation of the transition distribution function of a Markov process. *Annals of Mathematical Statistics*, *40*, 1386-1400.

Roussas, G. (1990). Nonparametric regression estimation under mixing conditions. *Stochastic Processes and their Applications*, *36*, 107-116.

Roussas, G. (1991a). Estimation of transition distribution function and its quantiles in Markov processes: strong consistency and asymptotic normality. In G. Roussas (Ed.), *Nonparametric Funcional Estimation and Related Topics* (Vol. 335, p. 443-462). Kluwer.

Roussas, G. (Ed.). (1991b). *Nonparametric Funcional Estimation and Related Topics* (Vol. 335). Kluwer.

Stute, W. (1986). On almost sure convergence of conditional empirical distribution functions. *Annals of Probability*, *14*(3), 891-901.

Tucker, H. (1967). *A Graduate Course in Probability.* New York and London: Academic press.

Watson, G. (1964). Smooth regression analysis. *Sankhya, A*, *26*, 359-372.

Yakowitz, S. (1979). Nonparametric estimation of Markov transition functions. *Annals of Statistics*, *7*, 671-679.

Yakowitz, S. (1985). Nonparametric density estimation, prediction and regression for Markov sequences. *Journal of the American Statistical Society*, *80*, 215-221.